Extraction of Biographical Information from Wikipedia Texts

Sérgio Soares

Instituto Superior Técnico, INESC-ID
Av. Professor Cavaco Silva, 2744-016 Porto Salvo, Portugal

Abstract. Documents with biographical information are frequently found on the Web, containing information useful for many different applications. In this paper, we address the task of automatically extracting biographical facts from textual documents published on the Web. We propose to segment documents into sequences of sentences, afterwards classifying each sentence as describing either a specific type of biographical fact, or some other case not related to biographical data. For classifying the sentences, we experimented with classification models based on the formalist of Naive Bayes, Support Vector Machines, Conditional Random Fields and voting protocols, using various sets of features for describing the sentences. Experimental results attest for the adequacy of the proposed approaches, showing an F_1 score of approximately 84% in the 2-class classification problem, 65% for 7-class classification problem and 59% for the 19-class classification problem.

1 Introduction

As the Web technology continues to thrive, a large number of documents containing biographical information are continuously generated and published online. Online newspapers, for instance, publish articles describing important facts or events related to the life of well-known individuals. Such Web documents describing biographical information often contain both meaningful biographical facts, as well as additional contents irrelevant to describe the person (e.g., detailed accounts of the person's actions). Although humans mostly manage to filter the desired information, manual inspection does not scale to very large document collections. It is our belief that, if relevant biographical information in human-generated texts can be extracted automatically, this information can be used in the production of structured biographical databases, capable of supporting many interesting studies in Humanities and related areas.

In the context of this Msc thesis, we addressed the information extraction problem of automatically extracting meaningful biographical facts, specifically immutable (e.g., place of birth) and mutable (e.g., occupation) personal characteristics, relations towards other persons and personal events, from textual documents, through an approach based on sentence classification.

We propose to first segment documents into sentences, afterwards classifying each sentence as describing either a specific type of biographical fact, or some

other case not related to biographical information. For classifying the sentences, we experimented with models based on the Naive Bayes, Support Vector Machines, Conditional Random Fields and voting protocols, using various sets of features for describing the sentences.

Experimental results attest for the adequacy of the proposed approaches, showing an F_1 score of approximately 84% in the 2-class classification problem when using a Naive Bayes classifier with token surface form, length, position and surrounding based features. The F_1 score for the 7-class classification problem was approximately 65% when using the Conditional Random Fields classifier with token surface, length, position, pattern and named entity features. Finally, the F_1 score for the 19-class classification problem was approximately 59% when using a classifier based on voting protocols with length, position pattern, named entity and surrounding features.

The rest of this paper is organized as follows: Section 2 presents related work. Section 3 presents the proposed taxonomy, the used corpus, the sentence classification methods and also details the features used for building the NB, SVMs and CRFs models. Section 4 presents the experimental validation of the proposed methods. Section 5 summarizes the results and highlights the verified tendencies. Finally, Section 6 presents our conclusions and points directions for future work.

2 Related work

The structuring of unstructured text has been studied by many authors in the field of automatic Information Extraction (IE), and it has also been extensively used in the real world applications [11]. The current trend in most IE tasks, including the extraction of biographical information from documents, is to use machine-learning approaches, relying on features extracted from training data that reflect properties of (i) individual tokens (e.g., capitalization, character types, or frequency) and (ii) general statistical measures either at a document scale, sentence scale, or corpus scale. Machine learning approaches are more attractive than rule-based methods in that they are trainable and adaptable, at the same time maintaining state-of-the-art performance. Several different classification methods have been successfully applied on information extraction, including Naive Bayes (NB), Support Vector Machines (SVMs), and Conditional Random Fields (CRFs) [6]. The method reported in this paper is based on the CRF tagger implementation available from the MinorThird toolkit[1], and on the NB and SVM implementations from the Weka toolkit[2].

Most information extraction tasks are modeled as sequence labeling problems where the objective is to label each token in a given sequence (i.e., the sequence of words in a document) as belonging to a specific class (e.g., *person*, *location*, *organization* or *other*). The main difference between our work and the main

[1] `http://sourceforge.net/apps/trac/minorthird/wiki`
[2] `http://www.cs.waikato.ac.nz/ml/weka/`

stream work on information extraction is that we employ sentences as the basic units, in the extraction task, instead of tokens.

In what concerns information extraction methods specifically dealing with biographical information, we have that the majority of previous works have modeled the production of biographies from textual documents as a summarization problem [3, 4, 12, 13]. Still, these previously proposed systems are basically extractive (i.e., the generation of biography summaries is seen as a problem of extracting the meaningful parts from the documents), combining information extraction with smoothing and merging techniques to improve coherence and reduce redundancy.

For instance, the system described by Cowie et al. automatically produces personal profiles (i.e., biographies) based on input queries, consisting of a chronologically ordered list of events with links to the source documents [4]. The system is designed as a pipeline of three modules, consisting of (i) an information retrieval stage where documents containing the query keywords are filtered, (ii) a summarization stage where relevant sentences are extracted, and (iii) a merging and output stage where the biographical summary is produced. Document sentences are scored and ranked according to the containment of the query terms, and the system also attempts to assign a date to each sentence using simple pattern matching.

Zhou et al. presented a similar system where a Naive Bayes model based on n-gram features is used to classify sentences as belonging to one of ten categories, namely bio (e.g., birth dates, death dates, and so on), fame factor, personality, personal, social, education, nationality, scandal, work, and others, which is simply the absence of a biographical category [13].

Garera et al. developed a system capable of extracting seven types of biographical facts (birth date, death date, birthplace, occupation nationality, gender and religion) through the use of six novel techniques, which exploited different classes of information [5]. Those techniques included partially-tethered contextual patterns, attributes of co-occurring entities, broad-context topical profiles, inter-attribute correlations, and also human age distribution with the objective of reducing the number of false positives.

Biadsy et al. constructed a system capable of producing biographies using summarization techniques based on multi-document sentence extraction [1]. This system selects the most important sentences for the target person using coreference techniques relative to the person's name. Then, those sentences are filtered using a binary classifier that retains only the biographical sentences. As an additional filter, to remove duplicate information, the sentences are clustered and only one representative sentence of each cluster is retained. Finally, the sentences are reordered using an SVMs regression model trained on biographies.

Kianmehr et al. performed a study in which SVMs were compared against neural networks in the task of text summarization. The features used in this study included: paragraph location, thematic word count, title word count, sentence length, presence of pronouns, and key phrases. The results of this study show that (i) SVMs performs better than neural networks for a small set of

features, (ii) increasing the number of layers for the neural networks does not increase the performance and (iii) SVMs and neural networks typically have the same accuracy level, although SVMs are faster.

In the context of his PhD thesis, Conway also addressed the problem of reliable identification of biographical sentences, exploring techniques that had been previously used for genre classification, leveraging on corpus derived n-grams and syntactic features [3].

The work reported in this paper can be seen as an advancement over these previous approaches, in the sense that it explores the usage of a sequence labeling models, i.e., CRFs, in the task of classifying sentences as belonging to biography-related categories.

3 Extracting biographical information from text

A biography can be defined as an account of the series of facts and events that make up a person's life. Different types of biographical facts include aspects related to immutable personal characteristics (i.e., date and place of birth, parenting information and date and place of death), mutable personal characteristics (i.e., education, occupation, residence and affiliation), relational personal characteristics (i.e., statements of involvement with other persons, including marital relationships, family relationships and indications of professional collaboration) and individual events (i.e., professional activities and personal events). The above classes were taken into consideration when developing the information extraction approach described in this paper.

3.1 The Taxonomy of Biographical Classes

It is important to note that biographical facts can be expressed in the form of complete sentences, besides phrases or single words, although here we only model the problem at the level of sentences. Thus, our task of automatically extracting biographical facts essentially refers to classifying sentences into one of two base categories, namely biographical and non biographical. Furthermore, sentences categorized as biographical are sub-categorized as (i) immutable personal characteristics, (ii) mutable personal characteristics, (iii) relational personal characteristics, (iv) individual events, and (v) other. Sentences categorized as immutable personal characteristics are further sub-categorized as (i.1) date and place of birth, (i.2) parenting information, or (i.3) date and place of death. Sentences categorized as mutable personal characteristics are also further sub-categorized, in this case as either (ii.1) education, (ii.2) occupation, (ii.3) residence and (ii.4) affiliation. Sentences categorized as relational personal characteristics are further sub-categorized as (iii.1) marital relationship, (iii.2) family relationships, and (iii.3) professional collaborations. Finally, sentences categorized as individual events are sub-categorized as either (iv.1) professional activity, (iv.2) and personal events. Figure 1 illustrates the hierarchy of classes that was considered.

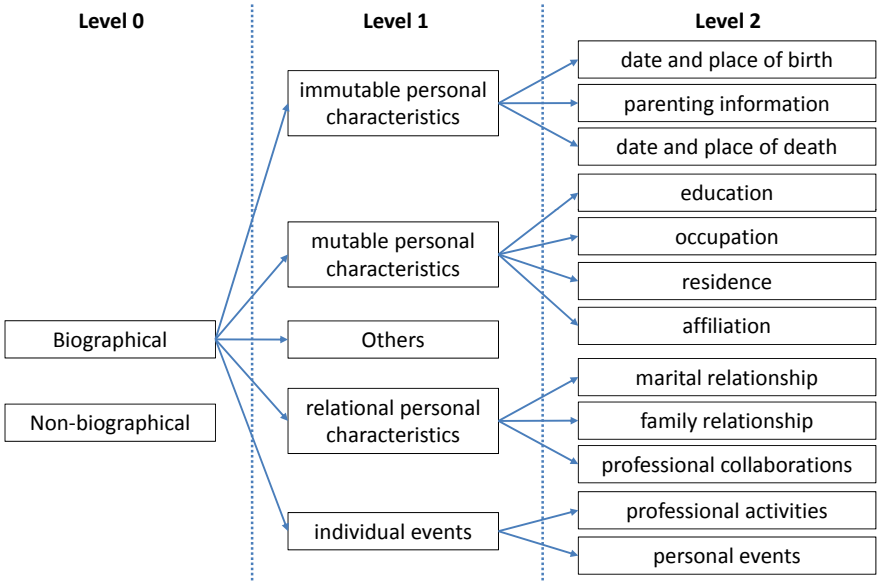

Fig. 1. The hierarchy of classes, considered in our tests.

Although the categories are hierarchical in nature, in this work, we will handle them in two distinct ways:

- As a simple list of 19 different categories without any hierarchy levels. Thus, when performing the classification, each sentence receives the most specific label which covers all the sentence's topics. For instance, if a sentence contains birth information, it is tagged as *date and place of birth*, but if the sentence also has information about the person's death, then the sentence will be labeled as *immutable personal characteristics*, because it is the most specific label that covers all the sentence's topics. Similarly, if a sentence has information about a person's education and professional collaborations, it will be tagged with the *biographical* label.
- As the three-level hierarchy that is presented in the Figure 1. When performing the classification with this hierarchy, the classification is done independently for each level. Thus, first we consider the level 0 which contains only the label *biographical* and *non biographical*. Consequently, the remaining labels are reduced to one of their accepted ancestors at the considered level (e.g., *date and place of birth* are reduced to *biographical*). This technique is used for each of the three hierarchy levels, but notice that for each level, the previous level's labels are also considered (e.g., when classifying the level 1 of the hierarchy, the label *biographical* existent in level 0 is still valid and do not need to be reduced).

Notice that when working in the last level of the hierarchy, all the labels are valid in a way similar to what happened with the flat working hierarchy mode, since all labels are allowed. However, in the hierarchical mode each sentence receives a top-level label (*biographical* or *non biographical*), and will only receive a more specific label (from level 1 and later for level 2) if the required conditions are met, contrasting with the flat hierarchy mode in which all the labels are considered to be on the same level and with the same probability. Finally, several different methods to traverse the hierarchy or either choose or not a more specific label, are possible, and those will be described later.

3.2 The Corpus of Biographical Documents

To build a corpus of gold-standard annotated data, we started by collecting a set of 100 Portuguese Wikipedia documents referring to football-related celebrities, like referees, players, coaches, etc. Afterwards, we performed the manual annotation of the documents, using the 19 different classes of the described taxonomy. Table 1 presents a statistical characterization of the resulting dataset.

	Value
Number of documents	100
Number of sentences	3408
Number of tokens	62968
Biographical	181
Immutable personal characteristics	4
Date or place of birth	2
Parenting information	20
Date or place of death	10
Mutable personal characteristics	4
Education	20
Occupation	212
Residence	6
Affiliation	5
Relational characteristics	1
Marital relationships	5
Family relationships	22
Professional collaborations	221
Individual events	78
Professional activities	484
Personal events	429
Others	221
Non biographical	1483

Table 1. Statistical characterization of the evaluation dataset.

3.3 The Proposed Methodology

Figure 2 provides a general overview on the methodology proposed for extracting biographical facts from textual documents. The first steps concern delimiting individual tokens over the documents, and also with segmenting the documents into sentences. In our case, this is done through the heuristic methods implemented in the LingPipe[3] text mining framework.

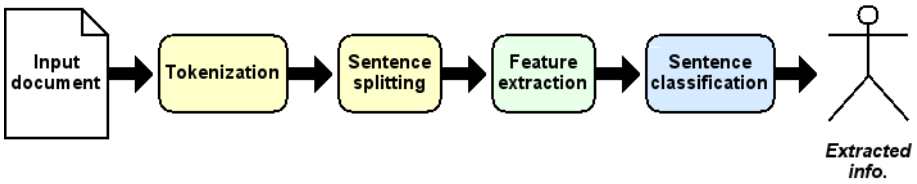

Fig. 2. The proposed extraction method for biographical sentences.

After delimiting the individual tokens and the sentences of the document, the next step is to extract a set of features describing each of the sentences. Subsection 3.5 details the considered features. The final step concerns classifying the sentences into one of the 19 classes mentioned above. The choosen classification models are described in the Section 3.4.

3.4 Classification Models

Classification approaches such as Naive Bayes or Support Vector Machines assume that the sentences in each document are independent of each other. However, in our particular application, the biographical facts display a strong sequential nature. For example, a place of birth is usually followed by parenting information, and documents with biographies usually present facts in a chronological ordering, starting from information describing a person's origins, followed by its professional achievements, and so on. Based on these observations, we also experimented with a sequence labeling model based on Conditional Random Fields, thus considering the inter-dependencies between sentences referring to biographical facts, in order to improve classification accuracy.

The remaining of this section introduces the theory behind the usage of NB, SVM and CRF models, in addition to the presentation of the features used to describe the sentences.

[3] `http://alias-i.com/lingpipe/`

Sentence classification with Naive Bayes Naive Bayes is a probabilistic model extensively used in text classification tasks. Naive Bayes classifiers base their operation on a naive independence assumption, considering that each feature is independent of the others. Thus, using estimations derived from a training set, it is possible to perform the classification by calculating the probability of each sentence belonging to each class, based on the sentence's features, and then choosing the class with the highest probability. The probability of assigning a class C to a given sentence represented by features $F_1, \ldots, F_N$ is calculated using the following equation:

$$P(C|F_1, \ldots, F_N) = \frac{P(C)P(F_1, \ldots, F_N|C)}{P(F_1, \ldots, F_N)} \quad (1)$$

In text classification applications, we are only interested in the numerator of the fraction, since the denominator does no depend on C and, consequently, it is constant. See [8] for a more detailed discussion of Naive Bayes classifiers.

Sentence classification with Support Vector Machines Support Vector Machines (SVM) are a very popular binary classification approach, based on the idea of constructing a hyperplane which separates the training instances belonging to each of two classes. SVMs maximize the separation margin between this hyperplane and the nearest training data points of any class. The larger the margin, the lower the generalization error of the classifier. SVMs can be used to classify both linearly and non-linearly separable data, through the usage of different kernel functions for mapping the original feature vector into a higher-dimensional feature space, they have been shown to outperform other popular classifiers such as neural networks, decision trees and K-nearest neighbor classifiers.

SVM classifiers can also be used in multi-class problems such as the one proposed in this paper, for instance by using a one-versus-all scheme in which we use a number of different binary classifiers equaling the number of classes, each one trained to distinguish the examples in a single class from the examples in all remaining classes [10]. The reader should refer to the survey paper by Moguerza and Muñoz for a more detailed description of SVM classifiers [9]. In this paper, we used an SVM implementation relying on the one-versus-all scheme and on a radial-basis function as the kernel.

Sentence classification with Conditional Random Fields The probabilistic model known as Conditional Random Fields offers an efficient and principled approach for addressing sequence classification problems such as the one presented in this paper [6].

We used the implementation from the MinorThird toolkit of first-order chain conditional random fields (CRFs), which are essentially undirected probabilistic graphical models (i.e., Markov networks) in which vertexes represent random variables and each edge represents a dependency between two variables.

A CRF model is discriminatively-trained to maximize the conditional probability of a set of hidden classes $y =< y_1, \ldots, y_C >$ given a set of input sentences $x =< x_1, \ldots, x_C >$. This conditional distribution has the following form:

$$p_\Lambda(y|x) = \frac{1}{\sum_y \prod_{c=1}^{C} \phi_c(y_c, y_{c+1}, x; \Lambda)} \prod_{c=1}^{C} \phi_c(y_c, y_{c+1}, x; \Lambda) \quad (2)$$

In the equation, ϕ_c are potential functions parameterized by Λ. Assuming ϕ_c factorizes a log-linear combination of arbitrary features computed over the subsequence c, then $\phi_c(y_c, y_{c+1}, x; \Lambda) = exp(\sum_k \lambda_k f_k(y_c, y_{c+1}, x))$ where f is a set of arbitrary feature functions over the input, each of which having an associate model parameter λ_k. The feature functions can informally be thought of as measurements on the input sequence that partially determine the likelihood of each possible value for y_c. The parameter k represents the number of considered features and the parameters $\Lambda = \{\lambda_k\}$ are a set of real-valued weights, typically estimated from labeled training data by maximizing the data likelihood function through stochastic gradient descent. Given a CRF model, finding the most probable sequence of hidden classes given some observations can be made through the Viterbi algorithm, a form of dynamic programming applicable to sequence classification models.

The modeling flexibility of CRFs permits the feature functions to be complex, overlapping features of the input, without requiring additional assumptions on their inter-dependencies. The list of features considered in our experiments is detailed in the following subsection.

3.5 The Considered features

In our experiments, we extracted various sets of features for building our NB, SVM and CRF models. The considered feature sets can be categorized as follows:

- Token features, including unigrams, bigrams and trigrams. The computation of these features starts by receiving a number N of n-grams to use, which was 500. Then, after listing all the unigrams, bigrams and trigrams from the training set, the application selects N of each. This selection is accomplished by removing repeatedly the most common and the most rare n-grams until the desired number of n-grams is obtained. After this filtering step, the remaining n-grams will be used as binary features for each sentence, where the corresponding value is one if it appears in the sentence and zero otherwise.
- Token surface features, referring to the visual features observable directly from the tokens composing the sentence (e.g., whether there is a capitalized word in the middle of the sentence, whether there is an abbreviation in the sentence, whether there is a number in the sentence, or whether the sentence ends with a question mark or an exclamation mark). We use one binary feature for each of the previously listed properties.
- Length based features, corresponding to the number of words in a sentence. Each possible sentence length is mapped to a binary feature which gets fired for a sentence having that particular length.

– Position based features, corresponding to the position of the current sentence in the document. We include this feature based on the observation that the first sentences in a document usually contain information about the origins of a person, while subsequent sentences often contain information regarding the person's activities. Thus, each sentence is mapped to a value between 1 and 4 according to its position in the text which could be from the first until the forth quarter, respectively.
– Pattern features, consisting in a list of common biographical words. This list includes expressions such as *was born*, *died* or *married*. To implement this feature, a stemmer for the Portuguese language is used to stem the sentence's words. Then, each stemmed word is compared with the stem of our list of biographical words. If a match exists, then the feature value is 1, else it is 0.
– Named entity features, referring to the occurrence of named entities of particular types (e.g., persons, locations and temporal expressions) in the sentence. Each named entity type is mapped to a binary feature which gets fired for a sentence having at least one named entity of that particular type. The generation of these features is based on a named entity tagger using the formalism of Hidden Markov Models, trained using a Portuguese dataset from Linguateca, known as HAREM[4].
– Surrounding sentence features, referring to the features observable from the previous and following sentence in the text. Specifically, these features refer to the token, token surface, position, length, domain specific and named entity features computed from the two surrounding sentences.

4 Evaluation Experiments

Our validation methodology involved the comparison of the classification results produced with different combinations of feature groups and with different classifiers, against gold-standard annotations provided by humans. In addition, we have created two new classifiers referred as *Voting 1* and *Voting 2* which use voting protocols to make their decisions. Voting protocols consists of an unsupervised approach, where each voter ranks the available candidates, and the outcome is chosen according to some voting rule [2]. The voting rule used in the context of this work is a variation of the rule known as *plurality rule*, in which each voter (NB, SVMs, CRFs) will vote in their preferred candidate, and the candidate with more votes is returned. Thus, both *Voting 1* and *Voting 2* use the tags returned by the voters (NB, SVMs, CRFs) to accomplish their own label assignment. The difference between *Voting 1* and *Voting 2* is that, when a voting draw occurs, the classifier *Voting 1* chooses the tag with biggest associated confidence, whereas the classifier *Voting 2* chooses the tag with more occurrences in the training data.

The used metrics were a macro-averaged precision, recall and F_1, weighted according to the proportion of examples of each class. The reason to use these metrics is the existence of large discrepancies between the number of examples

[4] `http://www.linguateca.pt/HAREM/`

available for each class. Thus, the use of these metrics avoids that the sentences with rare labels have a much higher weight than those with common labels.

The set of 100 hand-labeled documents was used for both training and validation, using five-fold cross validation for comparing the human-made annotations against those generated by the automatic tagger. In order to evaluate the impact of different feature sets, we divided the features presented in Section 3.5 into four groups, namely (i) token features and token surface features, (ii) length and position features, (iii) pattern and named entity features, and (iv) surrounding sentence features.

Finally, several types of experiments were conducted, namely:

- **Flat Classification** - This experiment considers that all the possible tags of the hierarchy are on the same level, ignoring the hierarchy levels. Thus, one of the existing 19 possible labels is selected by the classifiers based only on the training data.
- **Three Level Hierarchy** - This experiment used the taxonomy hierarchy described in section 3.1 with all its three levels. Consequently, for each experiment, three measurements are made (one for each level). Recall that for the measurement of each level, only the existent labels on that level plus their ancestors are accepted, and the remaining ones must be reduced to one of its accepted ancestors (i.e., if we are evaluating the level 1, all the *personal events* labels must be reduced to the *individual events* label). Thus, it is possible to measure the results in three different granularity levels. Furthermore, different techniques exploiting the referred hierarchy were tried, namely:
  - **Branching Method** - There are two distinct ways to select a new tag based on the hierarchical branches:
    - **Fixed Branch** - In the fixed branch classification mode, when the level of the hierarchy change, it is impossible to give a new tag which is not a descendant from the tag given in the previous level.
    - **Non Fixed Branch** - Is the opposite to the fixed branch classification mode, in which the new suggested label does not need to be a descendant of the previous assigned label.
  - **Increase Detail** - There are two distinct ways to go deeper in the tag hierarchy:
    - **Biggest Confidence** - Choose the tag with the biggest classifier's confidence. In this mode, after been given a tag to a sentence at any given level, this tag will only be replaced at the next hierarchy level if the new tag has more confidence than the tag of the previous level.
    - **Dynamic Threshold** - Choose the tag based on a dynamic threshold. With this method, the training data is used to determine the confidence threshold for each level and classifier, that should be used to accept new labels. Thus, even if a new label is suggested with bigger confidence than the previous label, it will only be accepted if the confidence of that classifier for the new label is bigger than the computed dynamic threshold for that classifier in that level.

5 Summary of the Experimental Results

From the described experiments, we concluded that the voting classifiers are appropriated for the task of biographical sentence extraction, specially the one that in case of a draw occurrence, chooses the tag with more occurrences in the training data (referred as *Voting 2*). Moreover, we concluded that both CRFs and SVMs had a similar performance, although CRFs is generally better. On the other hand, NB was capable of the top and worst results, since its behaviour is very sensitive to the features used. Thus, NB can be very useful for the task of biographical sentence classification if a good selection of features is done.

Relatively to the use of features, we concluded that the number of features used in an experience does not influence the performance obtained. Thus, it is preferable to use few carefully chosen than a lot of them chosen at random.

Although there is not a clear winner feature group, and the results greatly vary between experiments, we have noticed some patterns:

- When considering the finest coarseness level, the performance of NB is negatively affected when the feature group (iv) is used, specially in conjunction with the feature group (i). However, the opposite is true when considering the coarsest level of the hierarchy.
- For the NB classifier, the feature group (ii) and (iii) generally produced the best results when considering the most finest coarseness levels, but was one of the worst feature combinations when considering the most coarse level.
- The combination of feature groups (i) and (iii) produced the best results for the fixed hierarchy experiments when considering the SVMs. However, the reverse occurred in the non fixed hierarchy, in which the group (ii) or the group (ii) and (iv) was the best choice.
- For the SVMs classifier, the feature group (iii) was always the worst choice on the most coarseness level, while the group (ii) was a good choice.
- For the CRFs classifier, the combination of feature groups (i) and (ii) produced the top results in all the experiments, specially when combined with the group (iii) which yielded the best result in the finest hierarchy level.
- The worst feature group for the CRFs classifier was the group (iii) for the most coarse level of the hierarchy, and the group (ii) or group (iii) or the combination of groups (ii) and (iv) for the hierarchy levels one and two.
- Both the voting classifiers had very similar results, which was expected since its behaviors differed only in case of voting draws.
- For the classifiers based on the voting protocols, the combination of feature groups (i) and (ii) generally achieved good results, specially when combined with the group (iii) which yielded the best results for all hierarchy levels.
- The worst feature group for the voting classifiers was the group (iii) for the most coarse level of the hierarchy, and the feature groups (ii) or group (ii) and (iv) for the hierarchy level one and two (except for the non fixed branch and biggest confidence experiment).

Relatively to the tests which use the biggest confidence mode and the ones which use dynamic threshold mode, we noticed that the dynamic threshold is best for the finest coarse level (hierarchy level 2), but the biggest confidence mode is best for the experiments on the hierarchy level one. Moreover, when comparing the experiments which involved the fixed branch and the ones which used the non fixed branch mode, we observed that the results for the non fixed branch experiments achieved far better results. Finally, when comparing the results of the flat hierarchy experiment with the non fixed branch hierarchy and dynamic threshold experiment, we noticed that both had a similar performance, and the best one depends on the used feature groups.

The top individual results for each hierarchy level were:

- For the hierarchy level zero the best F_1 score was 83.71% when using the NB classifier with the feature groups (i), (ii) and (iv) with the non fixed branch hierarchy and dynamic threshold mode.
- For the hierarchy level one the best F_1 score was 65.29% when using the CRFs classifier with the feature groups (i), (ii) and (iii) with the non fixed branch hierarchy and dynamic threshold mode.
- For the hierarchy level two the best F_1 score was 58.74% when using the VOT2 classifier with the feature groups (ii), (iii) and (iv) with the non fixed branch hierarchy and dynamic threshold mode.

Despite the relatively good overall results, when analyzing the classification results for each individual class, we noticed that the number of classes with a zero F_1 score increased with the increase of the hierarchy level. Furthermore, there is a direct correlation between the number of examples in the training set for a given type of sentence, and the number of correct answers for that same type. Thus, we believe that the presented results could be improved if the available dataset was increased, specially for the finest hierarchy level, in which nine classes obtained a zero F_1 score.

6 Conclusions and future work

This paper, addressed the task of automatically extracting meaningful biographical facts, specifically immutable (e.g., place of birth) and mutable (e.g., occupation) personal characteristics, relations towards other persons and personal events, from textual documents published on the Web. We approached the task as a sentence classification problem, and we experimented with classification models based on the Naive Bayes (NB), Support Vector Machines (SVMs), Conditional Random Fields (CRFs), and two other classification methods, which used voting protocols, also using various sets of features for describing the sentences.

Distinct forms of classification were attempted also considering three different levels of coarseness, forming a hierarchy. Those forms of classification varied in the method used to traverse the hierarchy, and the method used to decide to choose or not, a finer grained class. From the experiments, we concluded that

the non fixed hierarchy branch mode, and the dynamic threshold mode produced the best results. However, the simplest classification experiment which used the flat hierarchy produced similar results.

Although many differences exist between this work and the Zhou et al. and Conway's work, we noticed that the best results, in the level zero of the hierarchy (biographical vs non biographical), was remarkably similar (83.71%, 82.42% and 80.66%), also confirming the results reported by other authors [7].

We also observed that the overall F_1 results decreased with the increase of the hierarchy level, which was expected since the number of possible classes increases as well. However, despite the good overall results, some individual classes obtained a zero F_1 score, due to the reduced number of examples, specially on the finer hierarchy level. Thus, we believe that the presented results could be improved if the available dataset was increased, specially for the finest hierarchy level, in which nine classes obtained a zero F_1 score.

The experiments allowed us to conclude that the classifiers based on voting protocols obtained generally the best results. However, there is not a clear classifier winner, since the results greatly varied from experience to experience. Similarly, there is not a winner feature combination. Consequently, the best combination will always depend on the objective of the classification. Furthermore, one can be interested on using a given combination to classify a given type of sentence even though that same combination produces bad overall results.

We think that classifying biographical sentences from textual documents still presents many challenges to the current state-of-the-art, specially when considering more than two classes. Thus, the returned results can hardly be used by other applications. Moreover, we believe that the existence of a common taxonomy and validation corpus would be extremely useful in order to compare different solutions, since actually the only thing in common between the most important works is the first level of the hierarchy (*biographical* and non biographical classes). Finally, it would be interesting to verify if the reported tendencies are portable to other areas not restricted to biographical information extraction and vice-versa, in order to incorporate new ideas in the field of biographical information extraction, from other information extraction areas.

Bibliography

[1] F. Biadsy, J. Hirschberg, and E. Filatova. An unsupervised approach to biography production using wikipedia. In *Proceedings of the Conference of Human Language Technologies of the Association for Computational Linguistics*, 2008.

[2] V. Conitzer. *Computational aspects of preference aggregation.* PhD thesis, Carnegie Mellon University, 2006.

[3] M. A. Conway. *Approaches to Automatic Biographical Sentence Classification: An Empirical Study.* PhD thesis, University of Sheffield, 2007.

[4] J. Cowie, S. Nirenburg, and H. Molino-Salgado. Generating personal profiles. Number Technical report, 2001.

[5] N. Garera and D. Yarowsky. Structural, transitive and latent models for biographic fact extraction. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, 2009.

[6] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, 2001.

[7] D. Lewis. *Representation and learning in information retrieval.* PhD thesis, 1992.

[8] A. McCallum and K. Nigam. A comparison of event models for naive bayes text classification. In *Proceeding of the AAAI/ICML-98 Workshop on Learning for Text Categorization*, 1998.

[9] J. M. Moguerza and A. Muñoz. Support vector machines with applications. *Statistical Science*, 21(3), 2006.

[10] R. Rifkin and A. Klautau. In defense of one-vs-all classification. *Journal of Machine Learning Research*, 5, 2004.

[11] S. Sarawagi. Information extraction. *Foundations and Trends in Databases*, 1, 2008.

[12] B. Schiffman, I. Mani, and K. J. Concepcion. Producing biographical summaries: Combining linguistic knowledge with corpus statistics. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, 2001.

[13] L. Zhou, T. M, and E. Hovy. Multi-document biographical summarization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2004.