Recognizing Biographical Sections in Wikipedia

Alessio Palmero Aprosio
Fondazione Bruno Kessler
Via Sommarive, 18
38123 Trento, Italy
aprosio@fbk.eu

Sara Tonelli
Fondazione Bruno Kessler
Via Sommarive, 18
38123 Trento, Italy
satonelli@fbk.eu

Abstract

Wikipedia is the largest collection of encyclopedic data ever written in the history of humanity. Thanks to its coverage and its availability in machine-readable format, it has become a primary resource for large-scale research in historical and cultural studies. In this work, we focus on the subset of pages describing persons, and we investigate the task of recognizing biographical sections from them: given a person's page, we identify the list of sections where information about her/his life is present. We model this as a sequence classification problem, and propose a supervised setting, in which the training data are acquired automatically. Besides, we show that six simple features extracted only from the section titles are very informative and yield good results well above a strong baseline.

1 Introduction

In the last years, several projects have started to address the mechanisms behind cultural development, borrowing techniques and algorithms from computer science and natural language processing to serve historical investigation. Efforts such as BiographyNet[1], Pantheon[2] or the Austrian Prosopographical Information System[3] prove an increasing interest in automatically extracting biographical descriptions from large amounts of data and combining them in a more general picture, taking advantage of the availability of such descriptions on the web. Wikipedia has been the main source of information for research in this direction despite its many biases, for instance its well-known English, Western and gender bias (Wikipedia Contributors, 2014). In fact, Wikipedia coverage both in terms of pages and in terms of languages, as well as the structured information that can be leveraged through DBpedia, has made it the primary resource for large-scale analyses on biographies. However, the lack of a consistent template for describing persons' lives led to the creation of a plethora of page types, where biographical information is displayed in diverse ways.

Based on a random sample of 100 persons' pages, we noticed that only 20% of them includes a section called *Biography* or *Life*, typically containing a set of subsections describing the main periods in a person's life from birth to death (see for instance https://en.wikipedia.org/wiki/Leonard_Bernstein). The other pages in our sample do not follow a pre-defined pattern and present the person's biography in one or several sections at the same level of the other ones (see for instance https://en.wikipedia.org/wiki/Judy_Holliday, with the *Filmography* and *Discography* sections at the same level of *Early Life* and *Career*). Given this high variability, it is very difficult to extract all and only those sections that describe a biography, and that build all together in sequence the description of a person's life. This depends also on the different types of non-biographical sections available, which in the case of prominent persons typically include main themes, reception, style, influences, legacy, work titles, etc. (see for instance *Will to power*, *Eternal return*, *Perspectivism*, *Critique on mass culture* in https://en.wikipedia.org/wiki/Friedrich_Nietzsche).

In this work, we present a simple methodology that, given a person's page in Wikipedia, recog-

[1] http://www.biographynet.nl

[2] http://pantheon.media.mit.edu/treemap/country_exports/IQ/all/-4000/2010/H15/pantheon

[3] http://www.oeaw.ac.at/acdh/de/node/188

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 811–816,
Lisbon, Portugal, 17-21 September 2015. ©2015 Association for Computational Linguistics.

nizes all sections that deal with his/her life even if no *Biography* section is present. The problem is modeled as a classification task using Conditional Random Fields, which are particularly suitable for our study because the biographical sections tend to follow a chronological order and present typical sequential patterns (for instance, the section *Early Life* is often followed by *Early Career*). While a simple token-based baseline is very difficult to beat when the task is performed at section level (i.e. deciding whether a section is biographical or not), our method performs best when the evaluation is performed at *page* level, recognizing *all* sections that describe a person's life. This is crucial if the task under investigation is meant as a preliminary step towards the automatic extraction of all events that compose a person's biography.

2 Related Work

To our knowledge, this is the first attempt to extract biographical sections from Wikipedia. Other past works focused on the recognition of biographical sentences (Biadsy et al., 2008; Zhou et al., 2004; Biryukov et al., 2005). However the two tasks have different goals: in our case, we aim at extracting all biographical sections, so that all events of a person's life from birth to death are present. The other approach, instead, is used to generate biography summaries, which was a task of the DUC2004 evaluation exercise[4]. Besides, while approaches for sentence selection look for textual features such as typical unigrams or bigrams that characterize biographical descriptions (Filatova and Prager, 2005), we adopt a much simpler approach by considering only section titles.

Other works focused on the analysis of typical events in selected articles from Wikipedia biographies by looking for a particular list of predefined events (Bamman and Smith, 2014). Our approach may complement such works by introducing a preprocessing step that extracts all and only the sections describing the biographies, upon which event extraction experiments can be performed. This would increase both the precision and the recall of the extracted information.

3 Experimental Setup

In this section we detail the data used for our experiments and the classification task.

3.1 Data set

Since our goal is to distinguish between biographical and non biographical sections, we focus only on Wikipedia pages describing persons. We derive our development, training and test data from the Pantheon data set (Yu et al., 2015), freely available for download.[5] The data set includes a list of 11,340 notable individuals with the link to their Wikipedia page in multiple languages, plus a number of additional information such as date and place of birth, category and language editions, which we do not consider for our study. Only the persons whose Wikipedia page is translated in at least 25 languages are included in Pantheon, as a proxy of prominent world personalities.

For each person in the list, we download the corresponding Wikipedia page in English and preprocess it using TheWikiMachine library[6]. Overall, we collect 11,075 pages, while 265 pages could not be retrieved because of problems with the links (mainly redirection links). We randomly select 100 pages as development set, 500 pages for test and the remaining 10,475 for building the training set.

For each page in the development and test set, we ask an annotator to assign a yes/no label to each section, to mark if it describes part of the person's life or not. We involve also a second annotator to label manually the development set (100 pages containing 834 sections). We compute Cohen's Kappa, which corresponds to 0.88. As a rule of thumb, this is considered an almost perfect agreement, which shows the clear-cut difference between biographical and non-biographical sections. The annotators use the Wikipedia original page to decide whether a section is biographical or not, therefore they can see both title and content of sections.

For the training set, we devise a novel methodology to acquire it completely automatically. We first extract from our collection of 10,475 pages the subset of pages containing a section called *Life* or *Biography*, which amount to 2,547. We consider such pages as a gold standard, since the presence of *Life* or *Biography* shows that their editors paid attention to the structure of the page, distinguishing between what belonged to the person's biography and what not. Therefore, all subsec-

[4] `http://duc.nist.gov/duc2004/`

[5] `http://thedata.harvard.edu/dvn/dv/pantheon`

[6] `https://bitbucket.org/fbk/twm-lib`

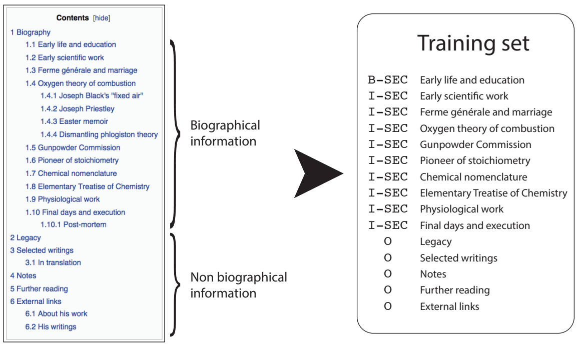
Contents [hide]
1 Biography
1.1 Early life and education
1.2 Early scientific work
1.3 Ferme générale and marriage
1.4 Oxygen theory of combustion
1.4.1 Joseph Black's "fixed air"
1.4.2 Joseph Priestley
1.4.3 Easter memoir
1.4.4 Dismantling phlogiston theory
1.5 Gunpowder Commission
1.6 Pioneer of stoichiometry
1.7 Chemical nomenclature
1.8 Elementary Treatise of Chemistry
1.9 Physiological work
1.10 Final days and execution
1.10.1 Post-mortem
2 Legacy
3 Selected writings
3.1 In translation
4 Notes
5 Further reading
6 External links
6.1 About his work
6.2 His writings
Biographical information
Non biographical information
Training set
B-SEC Early life and education
I-SEC Early scientific work
I-SEC Ferme générale and marriage
I-SEC Oxygen theory of combustion
I-SEC Gunpowder Commission
I-SEC Pioneer of stoichiometry
I-SEC Chemical nomenclature
I-SEC Elementary Treatise of Chemistry
I-SEC Physiological work
I-SEC Final days and execution
O Legacy
O Selected writings
O Notes
O Further reading
O External links

Figure 1: Extraction of training instances, from the Wikipedia page of Antoine Lavoisier (`https://en.wikipedia.org/wiki/Antoine_Lavoisier`). Sections are annotated in IOB2 format.

tions of *Life* or *Biography* are included as positive examples in the training set, because we assume that they all have biographical content. Instead, the remaining sections in the same pages represent our negative examples (see Figure 1). For the negative cases, we do not consider subsections because this level of detail is not needed for the classification, i.e. non-biographical sections contain only non-biographical subsections, and the classification of the first level is enough to propagate the corresponding label to the lower levels. Overall, the training set includes 2,547 sequences of sections (22,499 sections in total: 6,861 positive, 15,638 negative).

3.2 Classification experiment

We cast the problem as a supervised learning classification task, with the goal to label sequences of Wikipedia sections as describing a person's biography or not. As discussed in the introduction, we use Conditional Random Fields, since they are particularly suitable for sequence labelling (Lafferty et al., 2001). We use the implementation provided by CRFsuite[7] (Okazaki, 2007) for both training and classification tasks. The algorithm parameters are tuned on the development set.

In order to compare our approach with a non-sequential one, we perform the classification task also using Support Vector Machines (Vapnik, 1998). We use YAMCHA[8], a tool developed for chunking tasks, in which SVMs are easily combined with different context window sizes and dynamic features (Kudo and Matsumoto, 2001).

Since this work is only a preliminary step towards the automatic identification and extraction of biographical information from Wikipedia, we first experiment with the simplest approach. Therefore, we consider a small set of shallow features extracted *only from section titles*, and we ignore the content of the sections. Our six features are: the whole title, the tokens (lowercased), the bigrams, the first token of the section title, the first bigram, the position of the section with respect to the other sections in the same page (first, last, inside). We also use a sliding window of size 1 (both with CRF and SVM), so that the features extracted from the previous and the next sections are also considered to classify the current one. We experimented with window sizes > 1 on the development set, but they led to worse results.

Other features we implemented include the last word of the section, and a binary feature indicating whether the section title contains a year which is included between the date of birth and date of death of the person of interest. However, both pieces of information led to a performance drop on the development set, so we did not include them in the final feature set.

3.3 Baseline

To assess the performance of our system, we compare it with a baseline approach considering only the most frequent words in section titles. As shown in the evaluation (Section 4), this is indeed a very strong baseline. We identify the four most frequent tokens included in section titles, i.e. "Biography", "Life", "Death" and "Career", and we extract all sections that contain at least one of them (ignoring upper/lowercase). Finally, we consider as a positive example the smallest sequence of consecutive sections containing all of them. For instance, giving the sequence "Early life", "Career", "Presidency", "Death", "Legacy", "Awards", and "References", the first four sections are selected by the baseline as biographical sections: even if the "Presidency" word is not included in the tokens set, it is added as a consequence of its inclusion between "Early life" and "Death", both containing a token from the most frequent set.

[7] `http://www.chokkan.org/software/crfsuite/`

[8] `http://chasen.org/~taku/software/yamcha/`

4 Evaluation

We evaluate our system based on two different metrics, accounting for both exact and partial matches.

- In the *Exact* setting, a true positive is scored when *all* and *only* those sections with biographical information in a Wikipedia page are extracted. This measure is useful to understand how often it is possible to extract the *complete* and *exact* biographical text concerning a person.
- The *Intersection* measure, instead, assigns a score between 0 and 1 for every predicted sequence of sections based on how much it overlaps with the gold standard sequence (Johansson and Moschitti, 2013).

Evaluation results are reported in Table 1. The CRF-based approach outperforms the baseline in both configurations, with the highest improvement in the exact setting (+0.111 F1). Compared with the classification performance obtained with SVMs, CRFs yield better results only in the exact setting, while the intersection-based performance does not show substantial differences. In general, CRFs achieves a better precision but a lower recall than SVMs. If we look at the average length (in sections) of the false positive sequences, it is 3.72 for SVMs and 2.71 for CRFs. This difference confirms the behaviour of the SVM-based approach, which tends to overestimate the amount of biographical sections that should be tagged in sequence.

The results in Table 1 show also that a simple baseline relying on the four most frequent tokens in section titles achieves surprisingly good results, especially with the intersection-based metrics. This means that this basic approach tends to recognize correctly at least some of the sections describing a person's biography, because they present some recurrent patterns in their titles. In general, we observe that section titles alone are good indicators of their content, also without the need of more complex features. Although Wikipedia editors are free to name the sections and decide how to arrange them, there are some patterns that can be easily recognized automatically, especially by means of CRF.

	P	**R**	**F1**
CRFsuite			
Exact	0.694	0.662	0.677
Intersection	0.933	0.863	0.897
Yamcha SVM			
Exact	0.605	0.630	0.617
Intersection	0.857	0.942	0.898
Baseline			
Exact	0.584	0.548	0.566
Intersection	0.882	0.809	0.844

Table 1: Classification results using the *Exact* and *Intersection* settings

5 Discussion

We manually inspected the output of the classifier to identify possible issues. Apart from single classification mistakes, mainly due to unusual section titles that do not appear in the training set (such as "Anathematization" in Pope Honorius I page)[9], we found that some wrong classifications depended on specific types of persons in our data set. In particular, the classifier tends to assign a positive label to sections in the pages of mythological characters, even if they cannot have a biography because they did not exist. For instance, in the page of Apollo[10] there are sections entitled "Birth", "Youth", "Consorts and Children", which led the classifier to label them as biographical. Although mythological characters were included in the original Pantheon data set, we believe that they should be discarded, for instance by filtering them out a priori based on their Wikipedia categories.

Other false positives were found in pages of historical characters whose life was uncertain and was transmitted by others. For instance, the page of the geographer Pytheas[11] reports what the Roman author Pliny told about him, as well as what was said by other sources. These sections are very similar to those found in biographies, and are labelled as such.

We also performed additional experiments in order to investigate the impact of the size of the training data on the classification task. In particular, we extended our training data by creating new training sets based on 25,000, 50,000, 75,000 and 100,000 Wikipedia pages. These were obtained by

[9] `https://en.wikipedia.org/wiki/Pope_Honorius_I`

[10] `https://en.wikipedia.org/wiki/Apollo`

[11] `https://en.wikipedia.org/wiki/Pytheas`

ranking all pages with the PERSONDATA metadata according to the number of languages in which they are available, and then looking for the *Biography* and *Life* sections in the n top-ranked ones. Our evaluation shows that increasing the training size does not lead to a better performance, with an improvement of 0.01 with 100,000 pages over the results in Table 1 at the cost of a significant drop in processing speed.

6 Conclusions and Future Work

In this work we presented a simple yet effective approach to extract sequences of biographical sections from Wikipedia persons' pages. We model this task as a sequence classification problem using CRF and show that the section title alone conveys enough information to achieve a good classification result both in a supervised setting and with a rule-based baseline. Our contribution is threefold: *i)* we introduce the novel task of annotating the sequence of all sections describing a person's biography. This can be used as a preliminary step towards the extraction of all events characterizing a person's life; *ii)* we shed light on the regularities in Wikipedia persons' pages. Although Wikipedia is seen as a resource lacking consistency, with a flawed structure, our results show that at least persons' pages often present recurring patterns that are consistent across different biographies. The fact that only the most prominent figures have been included in the Pantheon data set is only a partial explanation of the good quality of such pages, because the English pages in our data set are often a reduced version of more extensive and edited pages in other languages. Finally, *iii)* we present an original approach to automatically acquire training data, using the pages with a *Biography* or *Life* section as gold data.

In the future, we plan to compare our approach based on section titles with more sophisticated approaches considering also the sections' content, to assess whether the latter improves over our simple methodology. Besides, we will build upon the outcome of this study by extracting the event sequence in a person's life starting from the complete biographies retrieved from Wikipedia.

The ongoing work is available as an open source project on GitHub[12] and is released under the GPLv3 license. In the project wiki one can find the dataset, the gold annotation and all the material needed to replicate the experiments.

Acknowledgments

The research leading to this paper was partially supported by the European Union's 7th Framework Programme via the NewsReader Project (ICT-316404). We would like to thank Naoaki Okazaki for the useful insight into CRFSuite.

References

David Bamman and Noah Smith. 2014. Unsupervised Discovery of Biographical Structure from Text. *Transactions of the Association for Computational Linguistics*, 2:363–376.

Fadi Biadsy, Julia Hirschberg, and Elena Filatova. 2008. An unsupervised approach to biography production using Wikipedia. In *In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL-2008).*

Maria Biryukov, Roxana Angheluta, and Marie-Francine Moens. 2005. Multidocument question answering text summarization using topic signatures. *Journal of Digital Information Management*, 3(1):27–33.

Elena Filatova and John M. Prager. 2005. Tell me what you do and i'll tell you what you are: Learning occupation-related activities for biographies. In *HLT/EMNLP 2005, Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 6-8 October 2005, Vancouver, British Columbia, Canada.*

Richard Johansson and Alessandro Moschitti. 2013. Relational features in fine-grained opinion analysis. *Computational Linguistics*, 39(3):473–509.

Taku Kudo and Yuji Matsumoto. 2001. Chunking with support vector machines. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics.*

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Naoaki Okazaki. 2007. CRFsuite: a fast implementation of Conditional Random Fields (CRFs). `http://www.chokkan.org/software/crfsuite/`.

Vladimir N. Vapnik. 1998. *Statistical Learning Theory*. Wiley-Interscience.

[12] `https://github.com/dkmfbk/biographies`

Wikipedia Contributors. 2014. Wikipedia: Systematic bias. https://en.wikipedia.org/wiki/Wikipedia:Systemic_bias. Accessed: 2015-06-14.

Amy Zhao Yu, Shahar Ronen, Kevin Zeng Hu, Tiffany Lu, and César A. Hidalgo. 2015. Pantheon: A Dataset for the Study of Global Cultural Production. *CoRR*, abs/1502.07310.

Liang Zhou, Miruna Ticrea, and Eduard H. Hovy. 2004. Multi-document biography summarization. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing , EMNLP 2004, A meeting of SIGDAT, a Special Interest Group of the ACL, held in conjunction with ACL 2004, 25-26 July 2004, Barcelona, Spain*, pages 434–441.