Information Sciences 465 (2018) 373–387

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

# Person name disambiguation on the web in a multilingual context

Agustín D. Delgado[a,*], Raquel Martínez[a], Soto Montalvo[b], Víctor Fresno[a]

[a] Department of Computer Systems and Languages, UNED Juan del Rosal, 16, Madrid 28040, Spain
[b] Department of Computer Science, Universidad Rey Juan Carlos Tulipán s/n, Móstoles 28933, Spain

ARTICLE INFO

*Article history:*
Received 20 February 2017
Revised 9 May 2018
Accepted 8 July 2018

*Keywords:*
Web people search
Multilingual clustering
Name disambiguation
Machine translation

ABSTRACT

Person Name Disambiguation on the Web is the problem of grouping web pages retrieved by a search engine when looking for a person name according to the individual they refer to. This problem has been addressed in a monolingual scenario where all the search results are written in the same language. However, search engines can also return links to web pages written in different languages. We study how to address multilingualism for this problem using the MC4WePS data set, a recent gold standard that includes real search results written in different languages. For this purpose, we first analyze the suitability of using a translation tool to treat multilingualism with two state-of-the-art clustering algorithms. Since the use of this kind of tools increases the processing time of the disambiguation process, we propose an approach to deal with multilingualism that generalizes the monolingual scenario and does not require any translation resources. Our approach obtains better results than the translation approaches with the gold standard, making it a competitive choice in a real scenario.



## 1. Introduction

Person Name Disambiguation problem (hereafter PND) has received the interest of Natural Language Processing (NLP), Information Retrieval (IR) and Information Extraction (IE) communities due to people names being a very ambiguous kind of Named Entities (NEs). This problem has been addressed in different scenarios. Since 2009, the Text Analysis Conferences[1] (TAC) have proposed the *entity linking* problem including some person names. The goal of this problem is to link mentions of an entity in a document to entities in a reference knowledge base, or to detect new entities. More recent works treat *author name disambiguation* [e.g. [17,23,27]], focusing on disambiguating researcher names, using as input a set of papers or references. Finally, PND has also been studied in the news domain [10]. However, our work is focused on PND in the web search scenario proposed by the WePS[2] (Web People Search) campaigns. This problem can be described as follows: given a query consisting of a person name in addition to the results retrieved by a search engine for that query, the goal is to cluster the resulting web pages according to the different individuals they refer to. Thus, the challenge lies in estimating the number of different individuals that share the same query name, and grouping the web pages that talk about the same individual

* Corresponding author.
*E-mail addresses:* agustin.delgado@lsi.uned.es (A.D. Delgado), raquel@lsi.uned.es (R. Martínez), soto.montalvo@urjc.es (S. Montalvo), vfresno@lsi.uned.es (V. Fresno).

[1] http://www.nist.gov/tac/tracks/index.html (Last access: 9 May 2018).
[2] http://nlp.uned.es/weps/ (Last access: 9 May 2018).

https://doi.org/10.1016/j.ins.2018.07.024
0020-0255/

in the same group. So, this problem can be formulated as a clustering problem where the number of resultant clusters is unknown. This scenario differs from *entity linking* because there is no knowledge base to link the different individuals, and it cannot be compared with *author name disambiguation* or PND in news, because in those problems the documents are of the same nature and contain specific structures and features that are often used but could not be found in web search results.

Internet users usually use search engines to look for people information. For instance, between 11 and 17% of web queries contain personal names [1] and four person names were in the top 10 of Google Search Trends for 2017[3]. Due to this fact, several start-ups (e.g. intelius.com, pipl.com, and so on) have launched search engines specializing in this kind of queries in recent years, but they are focused on social networking platforms. However, the most popular search engines (Google, Yahoo!, Bing, and so on) only provide disambiguation tools for celebrities or historical figures by means of the information obtained from their knowledge bases used to suggest entities related to the queries. For instance, the one employed by Google is called the *Knowledge Graph*[4], which uses information from several sources such as Wikipedia or the CIA World Factbook. Consequently, users have to look at the retrieved links, selecting and collecting those related to the individual they are interested in. In this situation, users usually refine the queries including additional terms, which could lead to missing other relevant web pages. For these reasons, it is still challenging to study PND methods in a web search scenario.

In a further complication, the heterogeneous nature of web results increases the difficulty of PND. For instance, some web pages related to a certain individual could be professional sites (e.g. corporation web pages), while others may contain personal information (e.g. blogs, social profiles, and so on) and both kinds of web pages could share very little common vocabulary. Particularly, Berendsen [7] concludes that the appearance of social networking platforms increases the difficulty of this problem. On the other hand, so far, PND systems have been evaluated for a monolingual scenario because the available corpora presented an evaluation framework where the query results were written in the same language. However, given a person name query, search engines can retrieve web pages written in different languages. For instance, web pages that contain professional information about a non-English native speaker might be written in English, while other web pages containing personal information about the same individual might be written in their native language. In addition, it is common to find search results written in different languages when looking for a person name shared by a celebrity. Therefore, multilingualism should be taken into account in a real search scenario, which makes this problem even more difficult. Finally, the computational cost of the disambiguation process should be as light as possible due to the real-time nature of PND on the Web.

In this paper, we study the impact of multilingualism on the PND problem on the Web. To this end, we use a recent gold standard that contains multilingual results. We first analyze the suitability of translating the search results by means of a machine translation tool using two state-of-the-art clustering algorithms to group the web pages. Next, we present an approach based on selecting suitable features when comparing web pages written in different languages that generalizes the monolingual scenario and avoids the use of translation resources.

The rest of the paper is organized as follows: Section 2 reviews work related to PND on the Web. Then, Section 3 presents the experimental framework, particularly, the data set and the clustering algorithms. Immediately afterwards, Section 4 presents a first approach for multilingual PND based on the use of a translation resource. Later, Section 5.1 details our approach for multilingual PND without translation. We discuss the results of our experimentation in Section 6. Finally, Section 7 presents some conclusions and future lines of work.

## 2. Related work

The PND problem proposed by WePS has been addressed in the state-of-the-art as a clustering problem. These PND systems are composed of two main phases: (1) web page representation, where the goal is to select suitable features from the web pages for this problem; and (2) applying a clustering algorithm to group web search results, so that each cluster contains all the web pages of a particular individual.

Regarding the representation of web pages, WePS evaluation campaigns showed that the participants mainly used the Vector Space Model (VSM). Some authors [5,18] explored topic modelling approaches, but obtained poor results. The most popular types of features used by WePS participants were: words, NEs and noun phrases, usually represented by a bag of words and weighted by means of TF-IDF. Nevertheless, some competitive systems [9,20,29] also enriched the representation by adding tokens extracted from the URLs, the titles of the web pages or the snippets returned by the search engine. Other works [5,9,29] applied attribute extraction techniques to obtain biographical information such as hints for distinguishing different individuals. Several works [13,25] concluded that $n$-grams co-occurrence between documents is positive evidence to decide if they refer to the same individual. This is due to the fact that the probability that $n$ random words appear in the same sequence in two different documents is very low when $n$ is high. Particularly, in Delgado et al. [13] we found that the $n$-grams composed by capitalized words are more useful for disambiguating person names. Finally, other works proposed the use of external resources in order to get additional features. For instance, Wikipedia could provide information on different individuals with the same name already disambiguated; Long and Shi [21] and Xu et al. [29] used Wikipedia entries for

[3] https://trends.google.com/trends/yis/2017/GLOBAL/ (Last access: 9 May 2018).

[4] https://www.google.com/intl/es419/insidesearch/features/search/knowledge.html (Last access: 9 May 2018).