Extraction of family relationships from historical documents

Julia Efremova*[1] and Toon Calders†[2]

[1]Eindhoven University of Technology, The Netherlands
[2]Université Libre de Bruxelles, Belgium

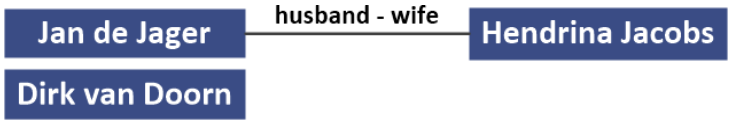
In this work we present an approach for the automatic extraction of family relationships from a real-world collection of historical notary acts. We retrieve relationships such as *husband - wife*, *parent - child*, *widow of*, etc. We study two ways to deal with this problem. In our first approach, we identify all person names in a document, generate all potential candidate pairs of names and predict whether they are related to each other using classification techniques where the text fragments that occur around and between two names are sued as features.

In the second approach, we train and apply a Hidden Markov Model to annotate every word in a document with an appropriate tag indicating if it is a name, a specified relationship descriptor, or neither of these. Then we look for the names connected to each other via relationship descriptors. We discuss the challenges such as processing raw data, obtaining a sufficient amount of training examples, and dealing with an imbalanced and noisy collection. We evaluate our results for each relationship type in terms of precision, recall and f - score.

Problem illustration. Below is an example of a notary act that has the *husband-wife* relationship (person names are underlined and relationships are in bold):

> Dit document certificeert: Jan de Jager en **zijn vrouw** Hendrina Jacobs, verklaren afstand te doen van alle rechten van de akte van koop en verkoop van 02/10/1906, opgemaakt voor notaris van Breda, ten behoeve van Dirk van Doorn, winkelier te Uden.
>
> *This document certifies: Jan de Jager and* ***his wife*** *Hendrina Jacobs, declare to waive all rights of the act of sale and purchase of 02/10/1906, registered at the notary Breda, with beneficiary Dirk van Doorn, shopkeeper in Uden.*

We see, that there are three persons mentioned in the document and the two of them have the *husband-wife* relationship:



People extraction from documents and prediction of their relationships are the main goals of this work. Details of applied techniques for family relationship extraction presented in research papers [2, 1].

References

[1] Julia Efremova, Alejandro Montes Garcia, Alfredo Jos Bolt Iriondo, and Toon Calders. Who are my ancestors? retrieving family relationships from historical texts. In *9th Summer School in Information Retrieval and Young Scientist Conference (RuSSIR'15)*, 2015.

[2] Julia Efremova, Alejandro Montes Garcia, Jianpeng Zhang, and Toon Calders. Towards population reconstruction: extraction of family relationships from historical documents. In *First International Workshop on Population Informatics for Big Data (21th ACM-SIGKDD PopInfo'15)*, 2015.

*i.efremova@tue.nl
†toon.calders@ulb.ac.be