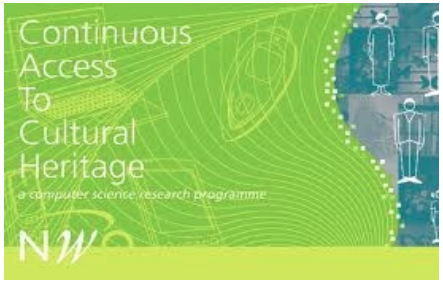


Extraction of family relationships from historical documents

Julia Efremova, Toon Calders

Co-authors

Toon Calders

Alejandro Montes García

Jianpeng Zhang

Collaboration:

BRABANTS HISTORISCH INFORMATIE CENTRUM

Introduction

Extraction of family relationships from historical documents

Content

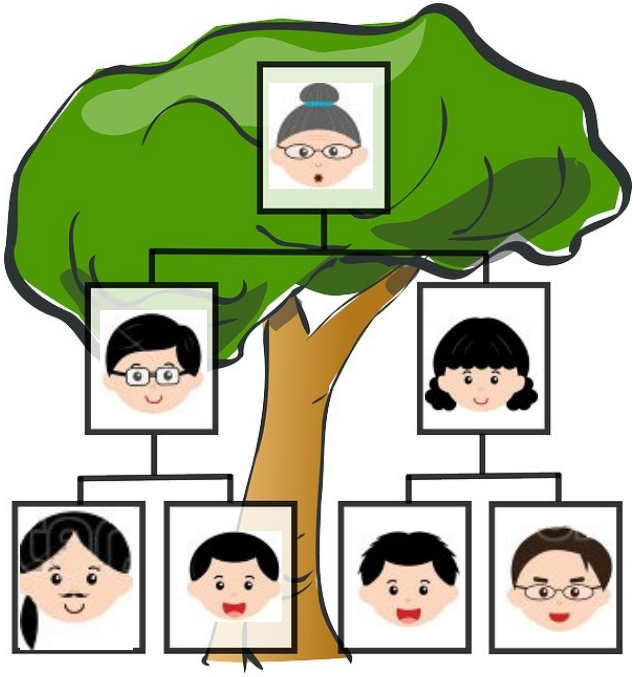
- Motivation and data description
- Data pre-processing
- Family relationship extraction
- Obtaining extra training data
- Experiments
- Conclusion & Future steps

Content

- **Motivation and data description**
- Data pre-processing
- Family relationship extraction
- Obtaining extra training data
- Experiments
- Conclusion & Future steps

Motivation

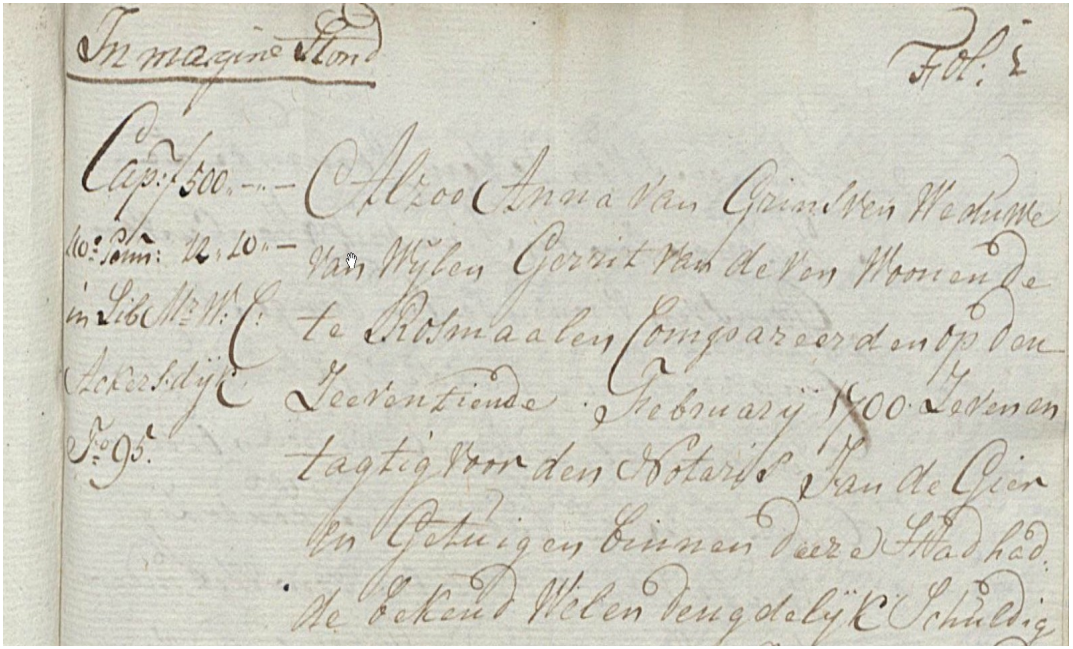
- Extracted family relationship are a part of a family tree
- Notary acts are a part of a family history



Sources of data

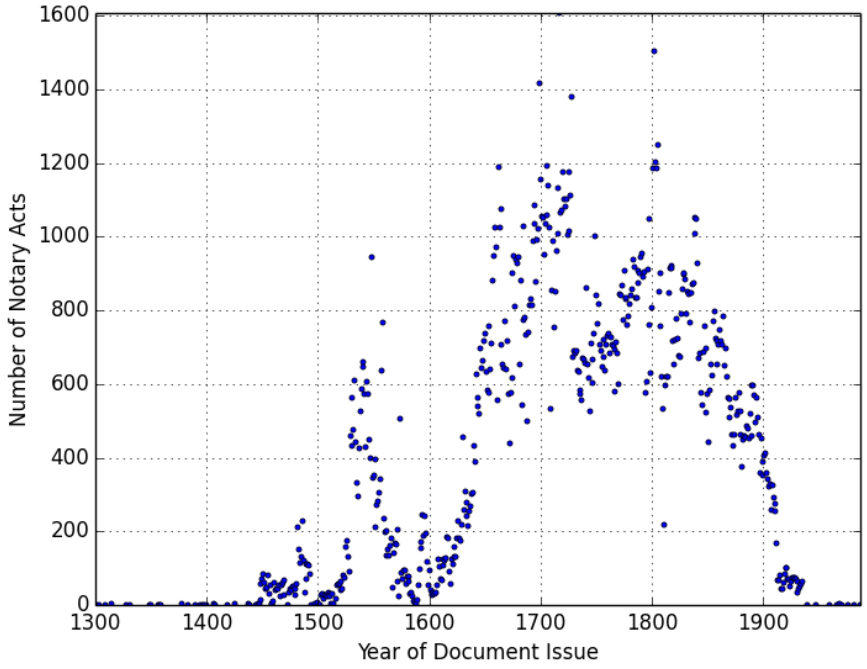
Archive data

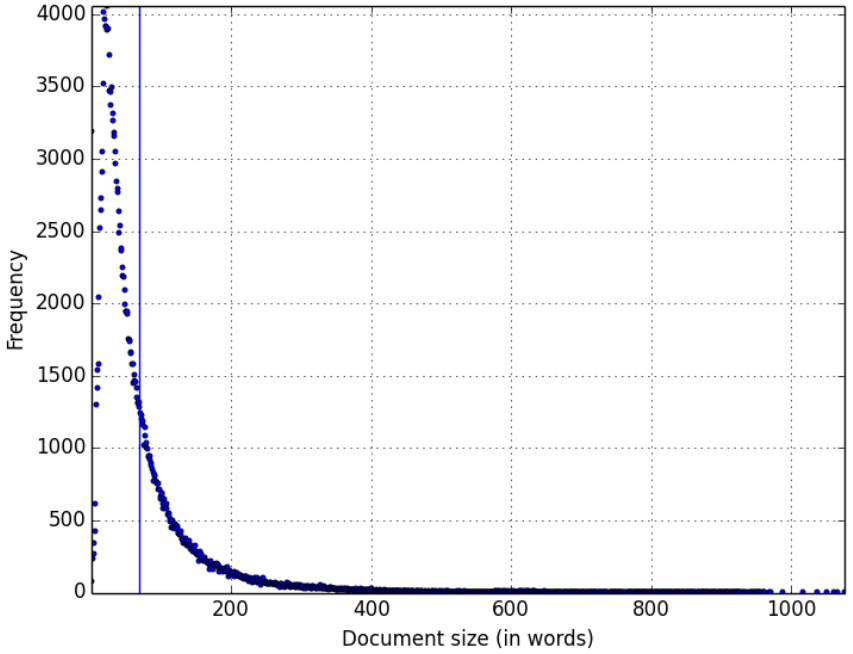
- **Historical notary acts**
- Criminal records
- Military records



Data Description

- Time period: 1400-1920
- Average length: 70 words
- ~ 115 000 documents in total

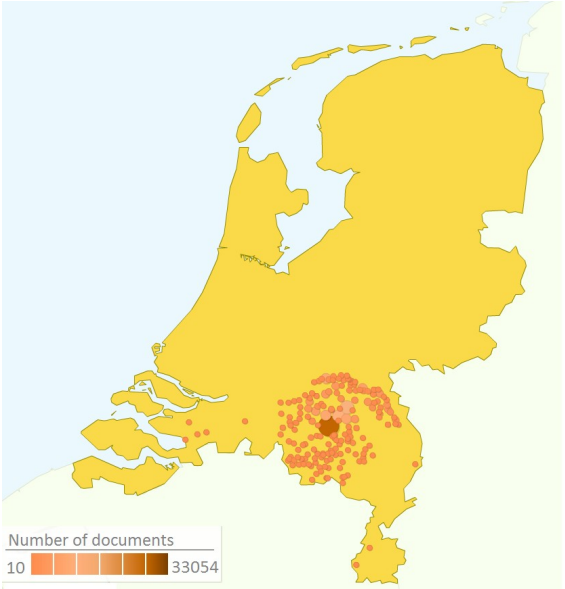




Main Categories

property transfer (transport), sale (verkoop), inheritance (testament), public sale of property (openvare verkoop), declaration (verklaring), partition of inheritance (erfdeling), resolution (resolutie)

aanbesteding aankondiging aansprakelijkstellingborgen aanstelling aflossing afstand akkoord belening belofte beslaglegging betaling bezwaarschrift borgstelling borgtocht brief certificatie dagvaarding dodenlijst eedsaflegging eenkindschap emancipatie erfdeling erfdienstbaarheid erfenis fundatie geldleningdoorcorpus handlichting herroeping huwelijk huwelijksevoorwaarden huwelijkstoestemming hypothecairelening ijking indorsatie inspectie inventaris jaarrente kwitantie lening machtiging meningsverschil nominatie obligatie ondervraging ontheffing ontlastbrief opdracht openbareverkoop openbareverpachting overeenkomst overzicht plakkaat poorterschap procuratie protest provisioneleverkoop rectificatie registratie remplacant rentebrief resolutie retroakte ruiling schenking schouw schuldbekentenis taxatie testament tochtrecht toestemming transport uitbesteding uitwinning verhuur verklaring verkoop vernadering verpachting verzegeling verzoek vest visitatie vonnis voogdij vrijwaring weddenschap wettiging wisselbrief zoenovereenkomst



An example of a notary act

- 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 Martinus 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 Martinus van Doorn, shopkeeper in Uden*.

Content

- Motivation and data description
- **Data pre-processing**
- Family relationship extraction
- Obtaining extra training data
- Experiments
- Conclusion & Future steps

Step 1: Data pre-processing

- Removing non-alphabetical symbols and stop words
- Extraction person names:
  - Own designed pattern-based name extraction
  - Frog tool (Dutch morpho-syntactic analyser)

Pattern-based name extraction

Why we need own name extraction?

- Low quality of data (old Dutch language)
- No available training data to train out-of the-box tool

Pattern-based name extraction

Available sources	Correspondent tag
First name dictionary (~ 46,000 first names)	<FN>
Last name dictionary (~115,000 last names)	<LN>

Additional information

Name prefix (van, de, …)	<P>
Initials	<I>
Start from capital letter	<CAP>

Pattern-based name extraction

- Jan de Jager

  Jan <FN> de <P> Jager <LN>

- Martinus van Doorn

  Martinus <FN> van <P> Doorn <CAP>

Name patterns:

- {<CAP>? <FN>+<CAP>? <I>? <P>? (<LN|CAP>)?}
- {<I>+ <FN>? <I>? (<LN|CAP>)+}
- {((<FN|CAP>)+ <P>)? <LN>}

Content

- Motivation and data description
- Data pre-processing
- **Family relationship extraction**
- Obtaining extra training data
- Experiments
- Conclusion & Future steps

Step 2: Family relationship extraction

Two general methods:

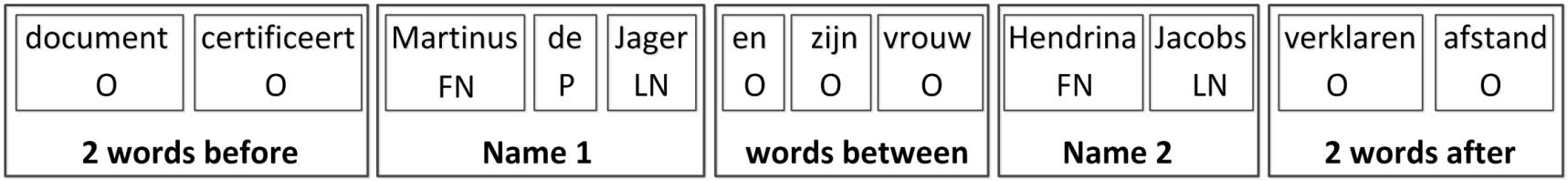
- Applying classification techniques
- Applying sequential data models

Classification approach

Family extraction process using classification approach

+ binary classification

Feature vector using Term Frequecy



HMM model for family relationship extraction

Family extraction process using HMM:

Annotation of **relationship descriptors** by HMM:

His <B-MAR> wife <I-MAR>

Husband <B-MAR> of <I-MAR>

HMM model for family relationship extraction

Applied Tags for HMM Annotation

Tag sets	Description
Person name annotation	{B-PER, I-PER, O}
Relation descriptors	{B-REL, I-REL, O}

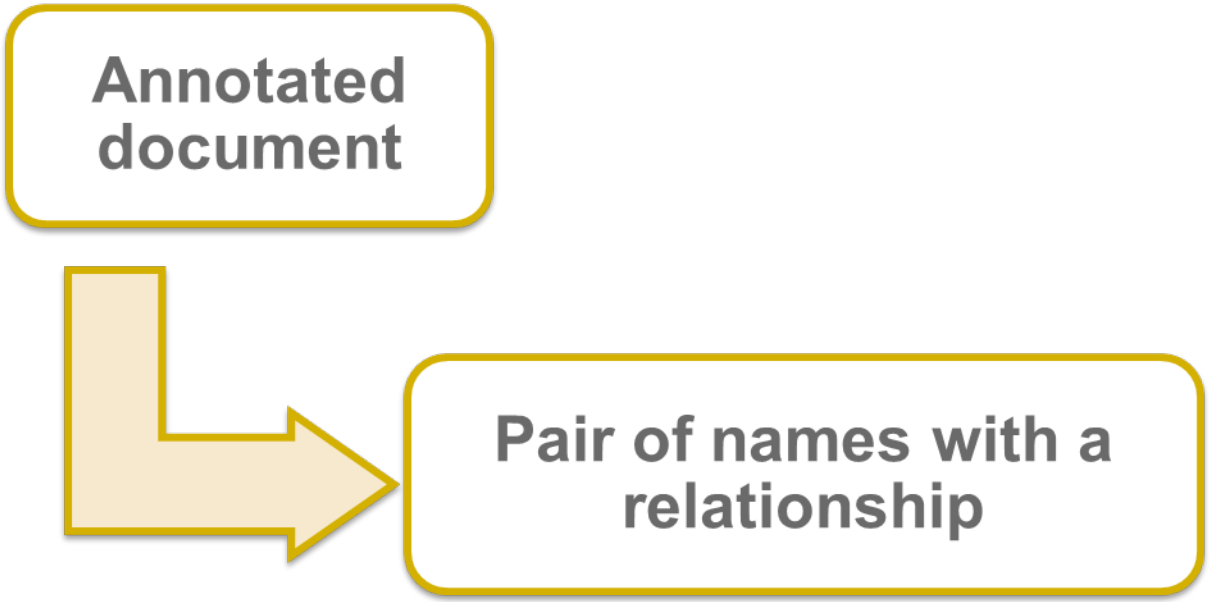
Jan **[B-PER]** de **[I-PER]** Jager **[I-PER]** and **[O]** his **[B-REL]** wife **[I-REL]** Hendrina **[B-PER]** Jacobs **[I-PER]**

HMM model for family relationship extraction

Typical family relationship:

- Marriage
- Parent of
- Widow of
- Sibling to
- Nephew of

Tag conversion and final pair generation



Conversion grammar:

- [PER, REL, PER]
- [PER]+`and'[PER]`,'[REL]

Content

- Motivation and data description
- Data pre-processing
- Family relationship extraction
- **Obtaining extra training data**
- Experiments
- Conclusion & Future steps

Obtaining extra training data

Frequent relationship descriptors:

Marriage	Parent	Widow of	Sibling to	Nephew	Auxiliary
married husband spouses	children child daughter	deceased died widow	sister brother sibling	nephew ant uncle	to, of, with from, his, her, their

Grammar of extra training data:

Family Relationship	Grammar
Marriage:	{<Au>?<M><Au>} {<Au><M><Au>?}
Parent-Child:	{<Au>?<P><Au>} {<Au><P><Au>?}
Widow of:	{<Au>?<W><Au>} {<Au><W><Au>?}

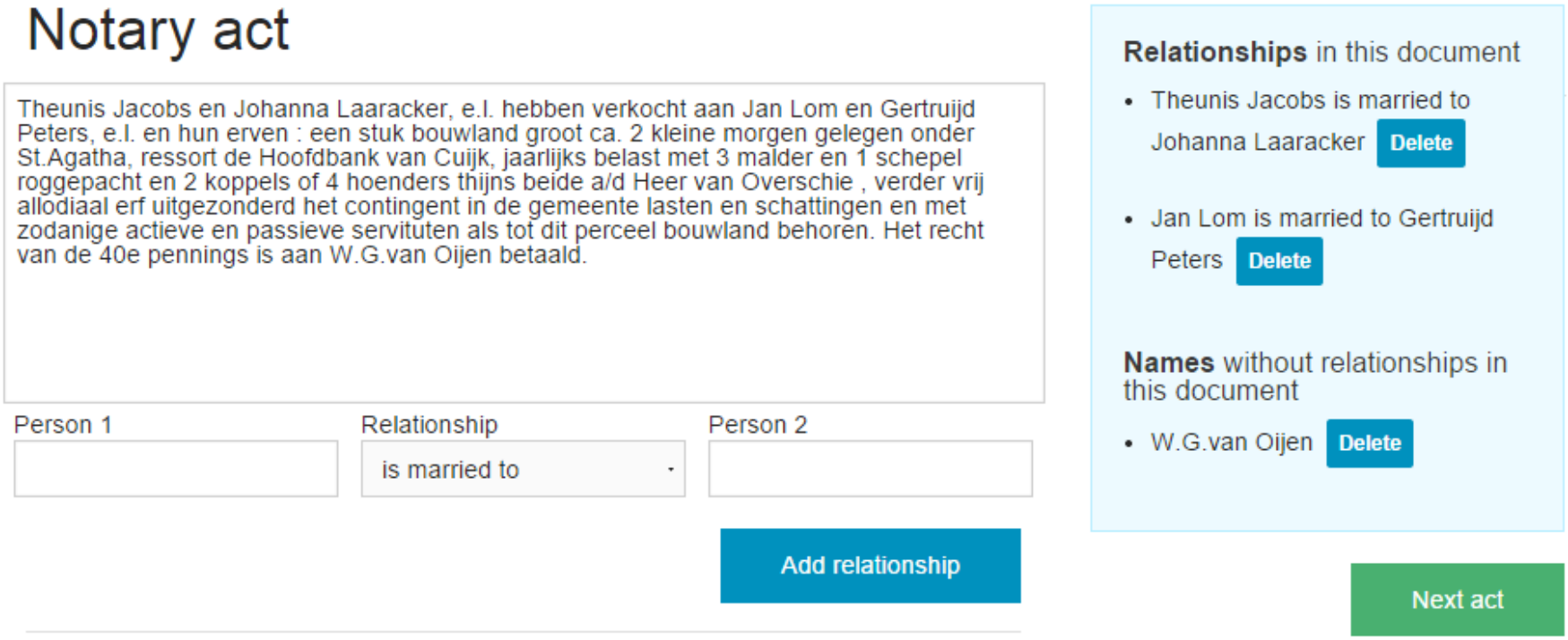
Content

- Motivation and data description
- Data pre-processing
- Family relationship extraction
- Obtaining extra training data
- **Experiments**
- Conclusion & Future steps

Experiments

- Manual labeling phase
- Learning model
- Cross validation

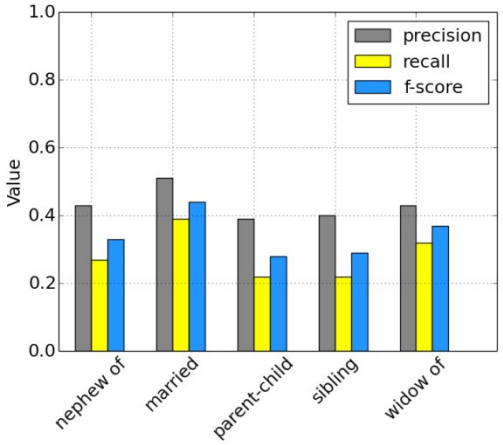
Labeling Tool



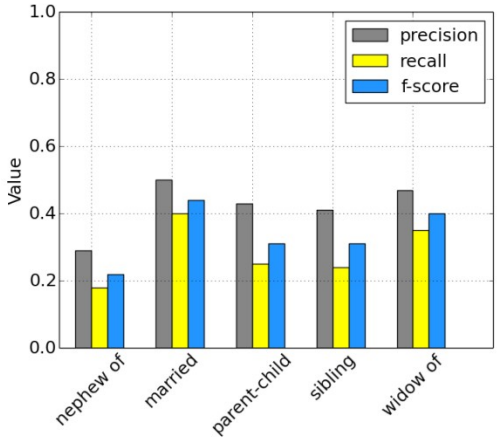
347 annotated notary acts
2000 annotated family relationships

Evaluation Results

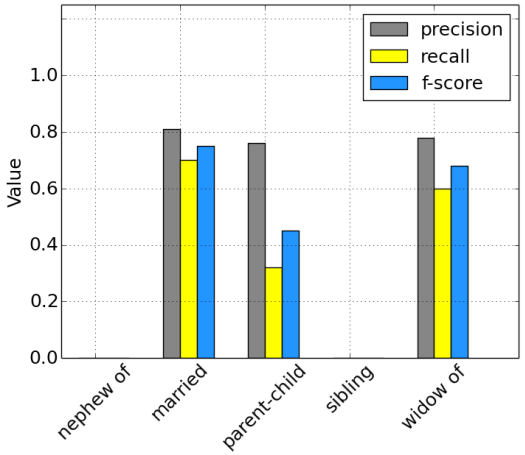
bi-grams standard classification



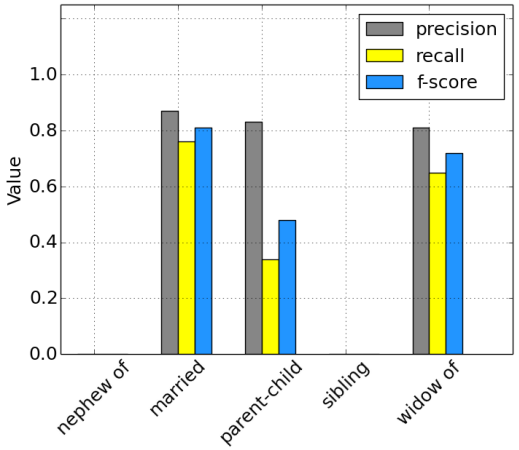
bi-grams and binary classification



HMM



HMM + NER



$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

Error analysis

Typical errors and reasons:

- Lack of representative training examples
- Overlapping pattern grammar (for HMM models)
- Implicit relationships

Content

- Motivation and data description
- Data pre-processing
- Family relationship extraction
- Obtaining extra training data
- Experiments
- **Conclusion & Future steps**

Conclusion

- A case study of family relationship extraction from historical documents
- Efficient methods suitable for a large data collection
- An important component of Genealogical research

Future Steps

- To combine approaches
- To deal with more efficiently with implicit relationships
- To build a family tree
- To reconstruct the history of every family
- To apply deep learning methods


Questions ?