Vocation Identification in Swedish Fiction

Dimitrios Kokkinakis[§], Ann Ighe[±], Mats Malm[#]
[§]Department of Swedish, Språkbanken; [±]Economic History, School of Business Economics and Law;
[#]Department of Literature, History of Ideas and Religion – University of Gothenburg, Sweden
{first.last}@gu.se

Abstract

This paper presents a system for automatic annotation of vocational signals in 19th century Swedish prose fiction. Besides vocation identification, the system assigns gender (male, female, unknown) to the vocation words. Since gender is a prominent attribute of *first names*, we apply a named-entity recognizer (NER) that uses first name gazetteers where each name has been pre-assigned gender, which aids gender assignment to vocations with unknown gender if appropriate context is available. We also use a statistical modelling method, conditional random fields (CRF), for learning gender-assigned vocations in combination with the results of the NER and other pattern matching techniques. The purpose of this work is to develop and apply tools to literature as means to expand our understanding of history in the area of literature-based gender studies, e.g. investigate how women enter literature, which functions do they assume and their working patterns. Vocation identification can be used as *one* such indicator for achieving some these goals.

1. Introduction

We present a system for the automatic annotation of vocational signals in Swedish text, namely 19th century prose fiction. *Vocation* in this context is used as a single or multi word expression term intended to capture the (professional) activities with which one occupies oneself, such as employment or other, wider, forms of productive occupations not necessarily paid. Therefore *vocation* is used here in a rather broad sense since we do not want to disallow word candidates that might not fit in a strict definition of the term.

Apart from vocation identification, the described system assigns gender, i.e. male, female or unknown, to the vocations by using various techniques. For instance, heuristics applied on hand-coded rules that use information based on personal pronouns, gender-bearing adjectives and gender-bearing suffixes. Since gender is a prominent attribute for a very large number of *first names*, we apply a named-entity recognition (NER) component that uses a first name gazetteer with 15,000 first names, in which each name has been pre-assigned gender. This way, if appropriate context is available, the NER can aid gender assignment of vocations with *unassigned* gender. Moreover, we also use conditional random fields (CRF; Lafferty et al., 2001), a statistical modelling method, i.e. for learning gender-assigned vocations in combination with the results of the rule-based system and the NER.

The purpose of this work is to use literature as means to expand our understanding of history by applying macroanalytic techniques in order to investigate how women enter literature as characters, which functions do they assume and their working patterns. The research questions themselves are not new, but in fact central to the field of gender studies. From a historical point of view, the 19th century in Sweden is a period with a dramatic restructuring of gender relations in formal institutions such as the civil law, and also a period where the separation of home and workplace came to redefine the spatial arenas for human interaction. Singular works of fiction have been analyzed in historical research but current development in digital humanities certainly opens new possibilities. Vocation identification can be used as one such indicator for achieving some these goals.

2. Background

During the last decade there has been a lot of research on applying automatic text analytic tools to annotate, enrich and mine historical material in various languages and for several reasons (Piotrowski, 2012; Rutner & Schonfeld, 2012; Jockers, 2013). The focus behind such research is to reduce the time consuming, manual work that is often carried out by historians and other literature scholars, in order to identify e.g. semantic associations, gender patterns and features or human networks (Agarwal et al, 2012). Sánchez-Marco et al. (2011) present an approach to part-of-speech that extends the coverage of a contemporary part-of-speech tagger to a historical variety of old Spanish by expanding the lexicon with historical forms. Pennacchiotti & Zanzotto (2008) investigated how modern Italian Natural Language Processing (NLP) tools, such as parsers, perform on historical texts. Various customization methods, such as manually building lexicons for different time periods or leveraging manually annotated corpora, showed performance improvements. Closer to our goals is the research by Pettersson & Nivre (2011); Fiebranz et al. (2011) and Pettersson et al. (2012), who in cooperation with historians, study what men and women did for a living in the Early Modern Swedish society (1550-1800).

3. Materials and Methods

The data we use in this work is a database of 18-19th century Swedish Prose Fiction – SPF – (for further details see here: <http://spf1800-1900>). The main part of this work involved the construction and adaptation of several lexical and semantic resources (tested) for modern Swedish to the language of SPF and algorithmic resources that use those for automatic labeling. As a starting point we used several hundreds of lexical units of relevant frames, such as

Medical_professionals and *People_by_origin*, from the Swedish FrameNet (<http://spraakbanken.gu.se/eng/swefn>). These lexical entries not only describe vocations but some are more general. Semi-automatically, all entries were assigned two features. The first was gender {Male, Female, Unknown}, based on e.g., typical gender bearing suffixes or head words in compound forms, and the second was *Vocation* or other related labels that describe various human qualities, such as *Performer*. All labels originate from the FrameNet. Moreover, we collected and structured other related terms, i.e. not only vocation but also other types of both generic and more specific ones that indicate person activities or relationship indicators, such as *Kinship*. The resulting lexicon contains over 18,000 terms and used for rule based pattern matching; 65% of all vocations in the lexical resources have been assigned *Unknown* gender. During processing we also use NER (only person identification) and post-NER pattern matching in order to assign gender to vocation words for which gender is marked as *Unknown*. First, a name gazetteer is used to assign gender to many vocation labels, for which their surface characteristics do not reveal gender; e.g., *bonden Petter* '(the) farmer Petter', here the NER will recognize *Petter* as human and male, information that will be propagated to the vocation word *farmer*, a term without encoded gender, which will get the same gender as its appositive *Petter*. We also use a complementary statistical modelling method, CRF, for learning gender-assigned vocations in combination with the results of the rule-based system and the NER (the vocation words together with basic features such as n-grams and word shape are used as features for the learner). For that purpose we use the Stanford CRF software (Finkel et al., 2005). A purpose of the CRF is for identifying vocations not captured by the previous techniques, which have very high precision. Training is based on a sample of 80,000 pre-annotated and manually inspected tokens (by the first author). The annotated sample was randomly selected from SPF; first automatically annotated by the rule based and NER components, and then sentences with at least one annotation were selected, manually inspected, corrected and used for training (60,000) and testing (20,000).

Since not all vocation annotations get gender assignment we use hand-coded rules for gender disambiguation. The heuristics applied in these rules include information on the usage of personal pronouns, gender-bearing adjectives and gender-bearing suffixes. E.g., personal pronouns such as *hans* 'his' and *hennes* 'her' are used for that purpose if they appear close to a vocation, e.g. *fiskaren och* ***hans*** *barn* 'the fisherman and his children', here *fiskaren* is identified as a vocation but with unknown gender which at this stage is assigned male since the pronoun *hans* is male and refers to the fisherman. Many older forms of Swedish adjectives were gender bearing; e.g. adjectives ending in *-e* designate male gender. For example, *fattig**e** bonden* 'the poor farmer', here *bonden* is identified as a vocation with unknown gender which is assigned male since the adjective *fattige* is designating a male head noun. Many noun suffixes or head words of compounds are gender bearing; e.g. suffixes *-erska* or *-inna* designate female gender; e.g. *tvätt**erska*** 'laundress' or *värd**inna*** 'hostess' are assigned female gender because of their gender bearing suffixes; gender bearing head words are also used for gender assignment; e.g. *bonde**fru*** [bonde+fru] 'peasant wife'.

4. Results and Discussion

The fact that a large number of vocations in the lexical resources have been assigned *Unknown* gender implies that processing requires to heavily relying on (wider) context to assign proper gender. This fact is mirrored on the results of e.g. the CRF evaluation, in which precision was 87.31%, recall 57.69% and f-score 69.47%. Different techniques have been tested, but still, a large number of vocation matches remains with unknown gender, so other more elaborated ways to identify gender using discourse information are required, e.g. CRF is used with default features, new/more features might be necessary to test.

Also, various problems could be identified during all stages. The most serious has been that a large number of vocations are assigned *Unknown* gender since there is no reliable context (at the sentence level) that can be used. Moreover, we have been restrictive to hard code gender for certain vocations in the resources, although, in principle, considering the nature of the texts, we could by default assign gender to a large number of these vocations. E.g., a number of military-related vocations, such as *löjtnant* 'lieutenant' are assigned *Unknown* gender in the lexicon, although these, predominantly, refer to males in the novels. Identical singular and plural forms of vocation terms are yet another problem, e.g. *politiker* 'politician' or 'politicians'. Part of speech annotation might have helped to eliminate the plural forms, but it is currently not used. Also, more elaborative models could be used to first determine who the personal pronouns refer to before an attempt could be made to assign the pronoun's gender to a vocation word with unknown one.

To conclude, in this work we have applied automatic text analytic techniques in order to identify vocation signals in 19th century prose fiction. Our goal has been to reduce the time consuming, manual work that is usually carried out by historians and other literature scholars, in order to e.g. identify gender patterns and semantic associations. For future work we would like to explore in more detail the variation of both the performance of the processing steps and also compare the results across time periods and authors' gender. Having as a starting point the described work we also want to investigate how is women's work and economic activities represented in prose fiction, and if there is a difference in how often male and female characters are mentioned with an occupational title and compare this longitudinally. Deeper analysis could provide interesting insights on the nature of which types of person activities are used by different authors, and thus confirm or reject established hypotheses about the kind of vocabulary used; e.g. do male authors use more vocation or kinship labels?

Acknowledgements

This work is partially supported by the Swedish Research Council's framework grant "Towards a knowledge-based culturomics" (<http://spraakbanken.gu.se/eng/culturomics>).

References

A. Agarwal, A. Corvalan, J. Jensen and O. Rambow. 2012. Social Network Analysis of Alice in Wonderland. Workshop on *Computational Lingusitics for Literature*. Pp 88–96, Montréal, Canada.

L. M. Jockers. 2013. *Macroanalysis - Digital Methods and Literary History. Topics in the Digital Humanities*. University of Illinois Press.

R. Fiebranz, E. Lindberg, J. Lindström and M. Ågren. 2011. Making verbs count: the research project 'Gender and Work' and its methodology. *Scan Economic History Review*. 59:3, 273-293.

J. R. Finkel, T. Grenager, and C. Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. Proceedings *of the 43rd Association for Computational Linguistics (ACL)*. Pp. 363-370.

J. Lafferty, A. McCallum, and F. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Proceedings 18th International Conf. on Machine Learning. Pp. 282–289. Morgan Kaufmann.

M. Pennacchiotti and F. M. Zanzotto. 2008. Natural Language Processing across time: an empirical investigation on Italian. *Advances in NLP*. LNCS, Vol 5221. Pp 371-382.

E. Pettersson and J. Nivre. 2011. Automatic Verb Extraction from Historical Swedish Texts. Proceedings of the *5th Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*. Portland, OR, USA. Pp 87-95.

E. Pettersson, B. Megyesi and J. Nivre. 2012. Parsing the Past – Identification of Verb Constructions in Historical Text. Proceedings of the *6th Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*. Avignon, France. Pp 65-74.

M. Piotrowski. 2012. *Natural Language Processing for Historical Texts*. Synthesis Lectures on Human Language Technologies. 5(2):1. Morgan & Claypool Publishers.

J. Rutner and R. C. Schonfeld. 2012. Supporting the Changing Research Practices of Historians. *National Endowment for the Humanities*. Ithaca S+R. New York.

C. Sánchez-Marco, G. Boleda and L. Padró. 2011. Extending the tool, or how to annotate historical language varieties. Proceedings of the *5th Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*. Portland, OR, USA. Pp 1-9.