Character Profiling in 19th Century Fiction

Dimitrios Kokkinakis
Center for Language Technology &
Språkbanken, Department of Swedish
University of Gothenburg, Sweden
dimitrios.kokkinakis@svenska.gu.se

Mats Malm
Department of Literature, History of
Ideas and Religion
University of Gothenburg, Sweden
mats.malm@lit.gu.se

Abstract

This paper describes the way in which personal relationships between main characters in 19th century Swedish prose fiction can be identified using information guided by named entities, provided by a entity recognition system adapted to the 19th century Swedish language characteristics. Interpersonal relation extraction is based on the context between two relevant, identified person entities. The relationships extraction process also utilizes the content of on-line available lexical semantic resources (suitable vocabularies) and fairly standard context matching methods that provide a basic mechanism for identifying a wealth of interpersonal relations. Such relations can hopefully aid the reader of a 19th-century Swedish literary work to better understand its content and plot, and get a bird's eye view on the landscape of the core story.

1 Introduction

Digitized information and the task of storing, generating and mining an ever greater volume of (textual) data become simpler and more efficient with every passing day. Along with this opportunity, however, comes a further challenge: to create the means whereby one can tap this great potentiality and engage it for the advancement of (scientific) understanding and knowledge mining. The goal of this research is to generate a complete profile for all main characters in each arbitrary volume in a literature collection of 19th century fiction. We also aim at a methodology that should be easily transferable to any other piece of literary work. A complete profile implies an exhaustive list of any kind of interpersonal relationships, such as *Friend Of* and *Antagonist Of* that can be encountered between the main characters in a literary work.

Similarly to social network extraction, there are numerous imaginable semantically oriented relationships between named entity pairs, this paper however only examines interpersonal ones. It also provides a brief description of the lexical resources and extended named entities, used by a Swedish named entity recognition (NER) system applied for the annotation of the collection. The NER system is to a great extent rule-based and uses a large set of lexically-driven resources. The system originates from a generic NER system used for annotation of modern Swedish which has been enhanced and improved by respecting common orthographic norms of nineteenth-century Swedish spelling.

One of the purposes in mind for this work is to test the applicability of Natural Language Processing (NLP) technologies in data from a deviant domain and time period than the ones the technology is designed for (i.e., contemporary, modern Swedish) in order to get a clearer picture of the strengths and weaknesses of the resources and tools and thus identify ways to further improve the obtained outcomes. This way we can facilitate the extraction of content-related semantic metadata, an important element in the management, dissemination and sustenance of digital repositories.

Name extraction in combination with filtering scripts that model the vocabularies, as well as fairly standard context matching methods provide a mechanism for identifying interpersonal relations that can also aid the reader of a literary work to better understand its content and plot, and get a bird's eye view on the landscape of the core story. Despite the risks of *spoilin*g the enjoyment that some readers of the narrative would otherwise have experienced without revealing of any plot elements, we still believe that such supporting aid can be used for an in-depth story understanding (for the human reader). Moreover, creating biographical sketches (e.g., birthplace)

Proceedings of Language Technologies for Digital Humanities and Cultural Heritage Workshop, pages 70–77,
Hissar, Bulgaria, 16 September 2011.

and extracting facts for entities (e.g., individuals) can be easily exploited in various possible ways by NLP technologies such as summarization and question answering (e.g., Jing *et al.*, 2007).

2 Related Work

Natural-language processing is an attractive approach to processing large text collections for relation extraction (usually defined as a relation predicate ranging over two arguments, e.g., concepts or people) and there exist a number of techniques that have applicability to any type of text; for a general review see Hachey (2009). Such techniques can facilitate more advanced research on literature and provide the appropriate mechanisms for generating multiple views on corpora and insights on people, places, and events in a large scale, through various types of relations.

Relation extraction was introduced in the mid 1990s by the *Template Element* and *Template Relation* tasks in MUC-6 (Message Understanding Conferences) and followed by the ACE (Automatic Content Extraction) *Relation Detection/Recognition* tasks (*cf.* Doddington *et al.*, 2004). Since then it has been an active and fruitful area of research, partly driven by the explosion of the available information via the Web and partly by the evidence that embedded relations are useful for various NLP tasks such as Q&A and Information Retrieval.

Relation extraction approaches (particularly binary ones) can be classified in various ways. Knowledge engineering approaches (e.g., rule-based, linguistic based), learning approaches (e.g., statistical, machine learning, bootstrapping) and hybrid ones; for an overview of techniques see Jinxiu (2007). Learning approaches become more and more common in the open domain i.e. large corpora of web scale, *cf.* Agichtein & Gravano, (2000); Christensen *et al.* (2010); relations are also of particular interest and prominent in the (bio)medical domain; e.g. Rosario & Hearst (2004); Giles & Wren (2008); Roberts *et al.* (2008). Elson *et al.* (2010) describe a method to extract social networks from literature (nineteenth-century British novels and serials) depending on the ability to determine when two characters are in conversation. The authors use a named-entity tagger to automatically locate all the names in a novel and then a classifier that automatically assigns a speaker to every instance of direct speech in the novel using features of the surrounding text. A "conversation" occurs if two characters speak within 300 words each other, and finally, a social network is constructed from the conversations. Nodes are named speakers and edges appear if there was a conversation between two characters, a heavier edge means more conversations. Our approach is mainly influenced by the work by Hasegawa *et al.* (2004) who proposed an unsupervised, domain-neutral approach to relation extraction by clustering named entity pairs according to the similarity of context words intervening between two entities and selecting the most frequent words from the context to label the relation.

3 Material: *a Prose Fiction Corpus*

Prose fiction is a just one type of textual material that has been brought into the electronic "life" using large scale digitized efforts. Prose fiction is an essential source within many disciplines of humanities (history, religion, sociology, linguistics etc) and social studies and an invaluable source for understanding the movements of society by its ability to demonstrate what forces and ideas are at work in the society of its time. Prose fiction is complex and difficult to use not only because of interpretational complexity but also because of its limited availability. The "19th Century Sweden in the Mirror of Prose Fiction" (*Det svenska 1800-talet speglat i prosafiktionen*) project (2009-12) aims at developing a representative corpus which mirrors society at given points in time, chronologically selected in such a way that historical comparisons can be made. The material is all fiction, written in the original and published separately for the first time, that appeared in Swedish during the years 1800, 1820, 1840, 1860, 1880 and 1900 (300 publications, ca 60,000 pages). The material provides a whole century of evolution and social, aesthetic, scientific, technical, cultural, religious and philosophical change.

3.1 Lexical Resources

The main focus of this research is the extraction of main character profiles[1], in literary archives and as a starting point we only look into interpersonal relationships. There is a number of suitable, freely available resources that we have started to exploit in order to aid the relation identification process, particularly the *RELATION-*

[1] Currently, this work is similar to the extraction of social networks but in the long run it is also desirable to extract more than merely interdependency relations of individuals (e.g., birth place, workplace etc.).

SHIP[2] vocabulary and two Swedish lexical semantic resources, namely the *FrameNet++*[3] and the *Swesaurus*[4]. These resources are useful and provide the appropriate machinery for our goals, namely to both identify appropriate relationship oriented lexical units and also appropriate relationship labels.

The RELATIONSHIP vocabulary defined by Davis & Vitiello (2010) is a good starting point for the labeling of the interpersonal relations. In their work Davis & Vitiello provide a description of 35 possible relationships that can occur between individuals. The description is not unproblematic since some of these relationships may be partially overlapping or even tautological such as *ChildOf* vs. *AncestorOf / DescedentOf* and *friendOf* vz. close*FriendOf*, The two other resources, namely the Swedish Swesaurus (Borin & Forsberg, 2010), that is fuzzy synsets in a WordNet-like resource under active development, and the Swedish FrameNet++ (Borin *et al.*, 2009) provides a large, and constantly growing number of synonyms and related words that are important for the relation extraction task.

In the Swedish FrameNet++ such words are called *lexical units* and are described by a number of *frames*. A frame is a script-like structure of concepts, which are linked to the meanings of linguistic units and associated with a specific event or state. A number of frames and particularly the lexical units encoded therein are relevant for interpersonal relationship extraction, such frames are for instance the *Personal_Relationship* (with lexical units: *flickvän* 'girl friend' and *make* 'husband', etc.), the *Kinship* (with lexical units: *barnbarn* 'grandchild', *bror* 'brother', *brorsdotter* 'niece', *dotter* 'daughter', etc.) and the *Forming_Relationship* (with lexical units: *förlova_sig* 'become engaged with', *gifta_sig* 'marry with', etc.). These frames are semi-automatically mapped to the RELATIONSHIP vocabulary and their containing lexical units become the actual lexical manifestation of the relationship in question. Similarly, we have experimented with the Swedish Swesaurus in order to identify synonyms for some of these lexical units. This way we can increase the amount of the words that can be part of various relations types. Thus, for the word *kollega* 'colleague' we can get a set of acceptable near synonyms such as *arbetskamrat* 'co-worker' but unfortunately also a number of not so suitable near synonyms such as *kompis* 'buddy', therefore we had to manually go through such near synonym lists and discard erroneous entries.

3.2 Named Entities and Animacy

There has been some work in the past on defining and applying rich name hierarchies, both specific (Fleischman & Hovy, 2002) and generic (Sekine, 2004) to various corpora. However, in other approaches (Kokkinakis, 2004) the wealth of name types is captured by implementing a fine-grained named entity taxonomy by keeping a small generic set of named entity types as *main* types and modeling the rest using a *subtype* mechanism. In this latter work a *Person entity* (a reference to a real word entity) is defined as proper nouns – personal names (forenames, surnames), animal/pet names, mythological names, names of Gods etc. – and common nouns and noun phrases denoting groups/sets of people. In this work the rule-based component for Person entity identification utilizes a large set of designator words (e.g., various types of nominal mentions) and phrases (e.g., typically verbal constructions) that require animate subjects, a relevant piece of knowledge which is explored for the annotation of animate instances in literary texts and other related tasks (*cf.* Orasan & Evans, 2001). These designators are divided into four groups according to their semantic denotation:

- nationality or the ethnic/racial group of a person (e.g. *tysken* 'the German [person]')
- profession (e.g. *läkaren* 'the doctor')
- family ties and relationships (e.g. *svärson* 'son in law'; *moster* 'aunt [from the mother's side]')
- individual that cannot be unambiguously categorized into any of the other three groups (e.g. *patienten* 'the patient')

Animacy markers are further marked for gender (male, female or unknown/unresolved such as *barn* 'child'). An example of animacy annotation is given below. In this example the animacy attribute, *ANI*, has a value *FAF* which stands for *FA*mily and *F*emale, while the attributes *TYPE* and *SuBT*ype refer to *PeRS*on and *HUM*an respectively: <ENAMEX TYPE="PRS" SBT="HUM" ANI="FAF">*Didriks mor*</ENAMEX> i.e. 'Didriks mother'. An important use of the animacy attribute is that it can be helpful for ruling out some erroneous non-allowable, gender-bearing

[2] http://vocab.org/relationship/.html
[3] http://spraakbanken.gu.se/eng/swefn
[4] http://spraakbanken.gu.se/swe/forskning/swefn/swesaurus

relations such as the one given in example 5. In this example there in an obvious *anomaly* involved considering the otherwise erroenous final relation, *SiblingOf*. *Stina* is recognized as female (through the attribute *UNF*) while in the preceding context the word *broder* 'brother' implies a following mention of a male gender, such features could be perhaps rule out spurious relationships.

4 Method

NLP techniques, such as Information Extraction, provide methods for identifying domain specific relations and event information. From a initial perspective such methods seem to be doomed to fail since each literary work is in itself a kind of *closed world* or domain where one may deal with death and resurrection and another on travelogues. However, each piece of work has certain general characteristics that can be captured by applying fairly standard NLP components such as named entity recognizers and indexers using various generic lexical resources, such as lexical units extracted from the Swesaurus. Also, inspired by similar methodologies that have shown high recall and precision figures, such as Hasegawa *et al.* (2004) we also try to capture interpersonal relationships by investigating ways in which the context between two entities can be modeled using unsupervised methods. Our basic approach is outlined below:

1. *Entity detection*: annotate corpora with named entities and animacy markers
2. *Context extraction*: extract sentences with co-occurring pairs of person named entities
3. *Relation detection and labeling:* label the extracted pairs of person entities
   a. automatically; for window size of 1-3 tokens using pattern matching templates with lexical units from the resources
   b. for window size of 4-10 tokens measure the context similarity between the extracted pairs of person entities
      i. make clusters of pairs and their context
      ii. semi-automatically label the clusters
4. *Merging:* filter, join and plot the results of (3a and 3b)

We automatically annotated each available volume in the collection with a slightly tuned to 19th century Swedish system; for details *cf.* Borin *et al.* (2007) and Borin & Kokkinakis (2010). We started by first clustering *all* possible context lengths and also applied template pattern matching once again on *all* possible contexts. After some experiments we split the process into two separate ones guided by the number of tokens between the entities. Very short contexts can be quickly and reliably captured in a pattern matching fashion. Therefore, we decided to *first* apply template pattern matching involving two recognized person entities and matched the intervened context with the lexical information extracted and modeled from the various lexical sources. Example of manually designed, pattern matching templates are provided below:

GrandparentOf: morfar|mormor|farfar|farmor|morfader|...

<PRSentity-1> *any?* {GrandparentOf} <PRSentity-2>
<PRSentity-1> {GrandparentOf} *any?* <PRSentity-2>
<PRSentity-1> *any?* <PRSentity-s-2> {GrandparentOf}

These template-examples attempt to capture the *GrandparentOf* relation by testing whether any of the lexical units, extracted from the resources as previously described, are between two person entities or immediately to the right of the second if this is in genitive form, i.e., ends in '-s'; *any* refers here to any optional non-empty sequence of characters while *GrandparentOf* is simply a convenient shorthand notation that gives a single name to a set of related lexical units. The results with this method were reliable when the intervening context is only a couple of tokens. The examples below illustrate the process; examples 1-3 contain the metadata obtained by the NER, and 1´-3´ the obtained relations after pattern matching and filtering.

(1) <ENAMEX TYPE="PRS" SBT="HUM" ANI="UNM">*Muhammeds*</ENAMEX> *dotter* <ENAMEX TYPE="PRS" SBT="HUM" ANI="FAF">*Fatima*</ENAMEX>; i.e., "Muhammeds daughter Fatima"

(2) <ENAMEX TYPE="PRS" SBT="HUM" ANI="UNU">*Strindberg*</ENAMEX> *hade träffat* <ENAMEX TYPE="PRS" SBT="HUM" ANI="UNF">*Nennie*</ENAMEX>; i.e., "Strindberg had met Nennie"

(3) <ENAMEX TYPE="PRS" SBT="HUM" ANI="UNF">*Taube*</ENAMEX> *anställde nu* <ENAMEX TYPE="PRS" SBT="HUM" ANI="UNF">*Marie Susanne Cederlöf*</ENAMEX>; i.e., "Taube employed now Marie Susanne Cederlöf"

(1′) *Muhammeds=>dotter/ParentOf=>Fatima*
ParentOf (Fatima, Muhammed)
(2′) *Strindberg=>träffat/HasMet=>Nennie*
HasMet (Strindberg, Nennie)
(3′) *Taube=>anställde/EmployerOf=>Marie Susanne Cederlöf*
EmployerOf (Taube, Marie Susanne Cederlöf)

For contexts between 4 and 10 tokens we produce context vectors (bag of words) from all intervening tokens of all contexts, with the exclusion of punctuation and numerical tokens. We chose not to include the very short contexts since pattern matching is reliable for short window sizes. After some test we limited the maximum window to be 10 tokens;, larger size of intervening tokens introduce in many cases noisy results. Examples 4-6 illustrate cases with a context between the person entities of >2 tokens. Note that the extracted relations in example 5′ is actually erroneous probably caused by one of the context words, namely *brodern* 'the brother'. This could be actually eliminated if the animacy attribute *ANI=UN**F** (**F**emale)* could be considered, a case left for future developments of this work. Example (6) illustrates another issue namely that of a potential relations that cannot be captured by the existing vocabulary; i.e. a tautology. For such relations there seems to be a *default* one, labeled, *Relationship* which is defined as "A class whose members are a particular type of connection existing between people related to or having dealings with each other", which we also use[5].

(4) <ENAMEX TYPE="PRS" SBT="HUM" ANI="UNF"> *Mafalda*</ENAMEX> *var van att se upp till syster* <ENAMEX TYPE ="PRS" SBT="HUM" ANI="FAF"> *Linda*</ENAMEX>; i.e., "Mafalda was used to seeing up to sister Linda"

(5) <ENAMEX TYPE="PRS" SBT="HUM" ANI="UNM"> *Ivar*</ENAMEX> *eggade med minspel och ögonkast brodern att trotsa, medan* <ENAMEX TYPE="PRS" SBT= "HUM" ANI="UNF">*Stina* </ENAMEX>; i.e., "Ivar edged with facial expressions and looks brother to defy, while Stina"

(6) <ENAMEX TYPE="PRS" SBT="HUM" ANI="UNU"> *Modén*</ENAMEX> *som av kamraterna också kallades* <ENAMEX TYPE="PRS" SBT="HUM" ANI="UNU"> *Moderat*</ENAMEX>; i.e., "Modén who of the colleagues also called Moderat"

[5] Here we could imagine an "Identical" relation since both names refer to the same individual. As a matter of fact we have recently initiated work in order to extend the list with missing relation types.

(4′) *Mafalda=>syster/SiblingOf => Linda*
SiblingOf (Mafalda, Linda)
(5′) **Ivar=>brodern/SiblingOf => Stina*
?SiblingOf (Ivar, Stina)
(6′) *Modén => kallades/Relationship => Moderat*
Relationship(Modén, Moderat))

5 Results

The context similarity between extracted pairs of entities can be measured in various ways. We applied hierarchical clustering (Seo & Shneiderman, 2002) with complete linkage and with cosine similarity as a similarity measure. We then manually evaluated obtained clusters and picked-up the top-5 most frequent words in these clusters as a means to characterize the cluster and tried to map these to the RELATIONSHIP vocabulary. Unfortunately, this activity revealed limitations since it was challenging to point to an appropriate label possibly because the data was *too* limited in size and also because most clusters had very few members (see more discussion below). For the evaluation (Precision, Recall and F-score) we chose to examine in more detail three randomly distinct volumes (see the References' section). *Precision (Pr)* is the fraction of relation instances that is correct, and for clustering Pr_{hc} the correct contexts that could be mapped among the contexts clustered automatically. *Recall (R)* is the fraction of relation instances that has been correctly extracted among all possible that involve two person named entities.

Table 1 summarizes these results, here *All* is the number of all window sizes with two person entities for a book and *<4* the number of contexts matched with the pattern matching approach. The abbreviated B1-B3 stand for the three volumes examined; B1 (Almqvist, 1847); B2 (Lo-Johansson, 1935) and B3 (Bergman, 1910).

	All	***<4***	$Pr_{<4}$	$R_{<4}$	$F_{<4}$	Pr_{hc}
B1	428	219	91.7% *24 rels*	70,6%	79,7%	47,1%
B2	227	115	93,7% *16 rels*	84,2%	88,7%	39,8%
B3	130	80	100% *9 rels*	75%	85,7%	41,8%

Table 1: Evaluation of relations in three books

$F_{<4}=2\times Pr\times R/Pr+R$

We manually inspected the <4 contexts and we found that only a small fraction of those (9) were wrong due to errors produced by the named entity tagger, e.g., <ENAMEX TYPE="PRS" SBT="HUM" ANI="UNM">*Kring*</ENAMEX> *sig hade* <ENAMEX

TYPE="PRS" SBT="HUM" ANI="UNU">*Kurt*</ENAMEX> *några...*; i.e., "Around him had Kurt some...", here *Kring* is simply an adverb..

6 Discussion

Our preliminary results showed that we need different strategies for modeling context between entities of interest and accordingly we have separated this modeling into two different relation detection methods, for short and longer context depending on the number of tokens intervening between named entities (within a single sentence). As one might expect, the use of patterns over very short contexts is much more successful than the clustering approach taken for long contexts. It seems that the unsupervised relation discovery approach is inappropriate to the application of extracting relationships from individual works of fiction. A problem with clustering such contexts is that one often gets a lot of small clusters and labeling is hard. Possibly because other work in the field or relation extraction generally assumes a very large corpus of (mainly) news texts, where relationships can be expected to be expressed multiple times in different documents and precision is improved through aggregation of mentions. However, in an individual work of fiction relationship are not expressed multiple times but rather once or twice. Therefore, this requires approaches with very high accuracy on individual relation mentions.

There is still another method that it would be lies in the gray zone between pattern matching and clustering. For instance Riloff (1996) applied more *generalised patterns* using regular expressions, e.g., *X* * *daughter* * *Y* where *X* and *Y* are person entities and * is any string of tokens, and she showed good results with this approach back in the 1990s.

The combined relations extracted can be viewed as a social network, i.e. a graph of relationships that indicate the important entities in a literary work and can be used to study or summarise interactions. The networks could also provide an alternative to standard presentation of information retrieval results when interacting with a litterary collection, e.g. by providing browsable representation of entities and their relationships that link to text passages where they are described.

Our first attempt to character profiling resulted in moderate precision and recall scores, at least for the clustering approach. But we believe that there is also plenty of scope for improvements and even of new research directions. For example, negations and speculative language might be a tricky issue since it can completely change the scope of a relation, e.g., *Han visste, att* <ENAMEX TYPE="PRS" SBT="HUM" ANI="UNF">*Mafalda* </ENAMEX> *icke tyckte om* <ENAMEX TYPE="PRS" SBT="HUM" ANI="UNM">*Zini*</ENAMEX> i.e., "He knew that Mafalda didn't like Zini". There are other issues that so far we have not confronted with, such as nameless characters, infrequently-appearing named characters, questions and opinions that once again can change the quality of a social network one experiences in a novel, e.g., *do you think X likes Y?*.

7 Conclusions

This paper has reported on initial experiments to automate character profiling in 19th century Swedish prose fiction. Profile implies intra-sentential relationship discovery between person entities. The aim is to support the users of digitized literature collections with tools that enable semantic search and browsing. In this sense, we can offer new ways for exploring the volumes of literary texts being made available through cultural heritage digitization projects.

In the future we also intend to even elaborate with relationships between main characters and other categories driven by named entities, such as between persons and locations and improve both the quantity and quality of the results. This way we can also extract significant properties of the characters and not only interpersonal relationships. It should be fairly straightforward since named entities can be reliably identified and a similar methodology as the one outlined in this paper can be applied. Applying other types of named entity types will eventually detect more relations about the characters and this will make the profiling more comprehensive than at the moment, which will reveal a clearer picture of the main characters' activities and associations. Another issue that needs attention is contexts with conjunctive mentions of entities, e.g. *X and Y*, since tokens in the near context might be good indicators of a relations as in the example: *Bröderne* <ENAMEX TYPE="PRS" SBT="HUM" ANI="UNM">*Tage*</ENAMEX> *och* <ENAMEX TYPE="PRS" SBT="HUM" ANI="UNM">*Robert*</ENAMEX>, i.e. "The brothers Tage and Robert".

At the moment we are looking at *explicit* relationships supported by textual evidence and did not include relations that dependent on the reader's understanding of the document's mean-

ing and/or her world knowledge, also a number of implicit relations could be inferred (e.g. *X ChildOf Y* implies *Y ParentOf X*). Moreover we would like to explore co-reference (pronominal references) since it plays an important role for profiling (biographical) extraction and for recognizing a larger set of relations between characters. Also *learning* of relationships in a complementary fashion in the future is envisaged and we plan to annotate data for this purpose.

Acknowledgments

This work is partially supported by the Centre for Language Technology (CLT) <http://clt.gu.se/> and the project *19th Century Sweden in the Mirror of Prose Fiction* financed by the "Research infrastructures" programme by the Swedish Research Council.

References

Carl Jonas Love Almqvist. 1847. *Herrarne på Ekolsund, del 1&2*. Samlade Verk 31. Almqvist & Wiksell International

Eugene Agichtein and Luis Gravano L. 2000. Snowball: Extracting Relations from large Plain-Text Collections. Proceedings of the 5th ACM International Conf. on Digital Libraries. New York.

Hjalmar Bergman. 1910. *Amourer*. Albert Bonniers Förlag, Stockholm.

Lars Borin, Dimitrios Kokkinakis and Leif-Jöran Olsson. 2007. Naming the Past: Named Entity and Animacy Recognition in 19th Century Swedish Literature. Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007), pages 1–8. Prague.

Lars Borin, Dana Dannélls, Markus Forsberg, Maria Toporowska Gronostaj and Dimitrios Kokkinakis. 2009. Thinking Green: Toward Swedish FrameNet++. Proceedings of the FrameNet Masterclass and Workshop. Milan, Italy.

Lars Borin and Markus Forsberg. 2010. Beyond the synset: Swesaurus – a fuzzy Swedish wordnet. Proceedings of the symposium: Re-thinking synonymy: semantic sameness and similarity in languages and their description. Helsinki, Finland.

Lars Borin and Dimitrios Kokkinakis. 2010. Literary Onomastics and Language Technology. In *Literary Education and Digital Learning. Methods and Technologies for Humanities Studies*. van Peer W., Zyngier S., Viana V. (eds). Pp. 53-78. IGI Global.

Janara Christensen, Mausam, Stephen Soderland and Oren Etzioni. 2010. Semantic Role Labeling for Open Information Extraction. Proceedings of the NAACL HLT First International Workshop on Formalisms and Methodology for Learning by Reading. Pp 52–60, Los Angeles, California.

Ian Davis and Eric Vitiello Jr E. 2010. RELATIONSHIP: A vocabulary for describing relationships between people. <http://vocab.org/relationship/.html>.

David K. Elson, Nicholas Dames, Kathleen R. McKeown. 2010. Extracting Social Networks from Literary Fiction. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010), Uppsala, Sweden.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel and Ralph Weischedel. 2004. The automatic content extraction (ACE) program tasks, data, and evaluation. Proc of the 4th Int Conf on Language Resources and Evaluation (LREC), pp 837–840, Lisbon.

Michael Fleischman and Eduard Hovy. 2002. Fine Grained Classification of Named Entities. Proceedings of the 19th International Conference on Computational Linguistics. Taipei, Taiwan. 1–7.

Cory B. Giles and Jonathan D. Wren. 2008. Large-scale directional relationship extraction and resolution. BMC Bioinformatics 2008, 9 (Suppl 9): S11doi:10.1186/1471-2105-9-S9-S11.

Benjamin Hachey 2009, *Towards Generic Relation Extraction*. PhD Thesis. Institute for Communicating and Collaborative Systems School of Informatics, University of Edinburgh.

Takaaki Hasegawa, Satoshi Sekine and Ralph Grishman. 2004. Discovering relations among named entities from large corpora. The 42nd Annual Meeting on Association for Computational Linguistics. Barcelona, Spain.

Hongyan Jing, Nanda Kambhatla, and Salim Roukos. (2007). Extracting social networks and biographical facts from conversational speech transcripts. Proceedings of the 45th Meeting of the Assoc. of Computational Linguistics, Prague, Czech Rep.

Chen Jinxiu. 2007. *Automatic relation extraction among named entities from text contents*. PhD thesis, University of Singapore.

Ivar Lo-Johansson. 1935. *Kungsgatan*. Albert Bonniers Förlag, Stockholm.

Dimitrios Kokkinakis. 2004 Reducing the Effect of Name Explosion. LREC Workshop: Beyond Named Entity Recognition, Semantic Labelling for NLP tasks. 4th LREC. Lissabon, Portugal.

Constantin Orasan and Richard Evans. 2001. Learning to Identify Animate References. Proceedings of the Workshop on Computational Natural Language Learning (CoNLL-2001). Toulouse, France.

Ellen Riloff. 1996. Automatically Generating Extraction Patterns form Untagged Text. Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI-96), pp. 1044-1049.

Angus Roberts, Robert Gaizauskas and Mark Hepple. 2008. Extracting Clinical Relationships from Patient Narratives. BioNLP 2008. Pp 10–18, Ohio, USA.

Barbara Rosario and Marti A. Hearst. 2004. Classifying Semantic relations in Bioscience Texts. Proceedings of the 42nd Annual Meeting on ACL. Barcelona, Spain.

Satoshi Sekine. 2004. Definition, Dictionaries and Tagger for Extended Named Entity Hierarchy. Proceedings of the Language Resources and Evaluation Conf (LREC). Lisbon, Portugal.

Jinwook Seo and Ben Shneiderman, 2002. Interactively Exploring Hierarchical Clustering Results. IEEE Computer, Vol. 35:7, pp. 80-86. <http://www.cs.umd.edu/ hcil/hce/>