A Link to the Past: Constructing Historical Social Networks

Matje van de Camp
Tilburg Centre for Cognition
and Communication
Tilburg University, The Netherlands
M.M.v.d.Camp@uvt.nl

Antal van den Bosch
Tilburg Centre for Cognition
and Communication
Tilburg University, The Netherlands
Antal.vdnBosch@uvt.nl

Abstract

To assist in the research of social networks in history, we develop machine-learning-based tools for the identification and classification of personal relationships. Our case study focuses on the Dutch social movement between 1870 and 1940, and is based on biographical texts describing the lives of notable people in this movement. We treat the identification and the labeling of relations between two persons into positive, neutral, and negative both as a sequence of two tasks and as a single task. We observe that our machine-learning classifiers, support vector machines, produce better generalization performance on the single task. We show how a complete social network can be built from these classifications, and provide a qualitative analysis of the induced network using expert judgements on samples of the network.

1 Introduction

The rapid growth of Social Networking Services such as Facebook, Myspace and Twitter over the last few years has made it possible to gather data on human interactions on a large scale, causing an increased interest in the field of Social Network Analysis and Extraction. Although we are now more interconnected than ever before due to technological advances, social networks have always been a vital part of human existence. They are prerequisite to the distribution of knowledge and beliefs among people and to the formation of larger entities such as organizations and communities. By applying the technology of today to the heritage of our past, it may be possible to uncover yet unknown patterns and provide a better insight into our society's development.

In this paper we present a case study based on historical biographical information, so-called secondary historical sources, describing people in a particular domain, region and time frame: the Dutch social movement between the mid-19th and mid-20th century. "Social movement" refers to the social-political-economical complex of ideologies, worker's unions, political organizations, and art movements that arose from the ideas of Karl Marx (1818–1883) and followers. In the Netherlands, a network of persons unfolded over time with leader figures such as Ferdinand Domela Nieuwenhuis (1846–1919) and Pieter Jelles Troelstra (1860–1930). Although this network is implicit in all the primary and secondary historical writings documenting the period, and partly explicit in the minds of experts studying the domain, there is no explicitly modeled social network of this group of persons. Yet, it would potentially benefit further research in social history to have this in the form of a computational model.

In our study we focus on detecting and labeling relations between two persons, where one of the persons, A, is the topic of a biographical article, and the other person, B, is mentioned in that article. The genre of biographical articles allows us to assume that person A is topical throughout the text. What remains is to determine whether the mention of person B signifies a relation between A and B, and if so, whether the relation in the direction of A to B can be labeled as positive, neutral, or negative. Many more fine-grained labels are possible (as discussed later in

Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, ACL-HLT 2011, pages 61–69,
24 June, 2011, Portland, Oregon, USA ©2011 Association for Computational Linguistics

the paper), but the primary aim of our case study is to build a basic network out of robustly recognized person-to-person relations at the highest possible accuracy. As our data only consists of several hundreds of articles describing an amount of people of roughly the same order of magnitude, we are facing data sparsity, and thus are limited in the granularity of the labels we wish to predict.

This paper is structured as follows. After a brief survey of related research in Section 2, we describe our method of research, our data, and our annotation scheme in Section 3. In Section 4 we describe how we implement relation detection and classification as supervised machine learning tasks. The outcomes of the experiments on our data are provided in Section 5. We discuss our findings, formulate conclusions, and identify points for future research in Section 6.

2 Related Research

Our research combines Social Network Extraction and Sentiment Analysis. We briefly review related research in both areas.

2.1 Social Network Extraction

A widely used method for determining the relatedness of two entities was first introduced by Kautz et al (1997). They compute the relatedness between two entities by normalizing their co-occurrence count on the Web with their individual hit counts using the Jaccard coefficient. If the coefficient reaches a certain threshold, the entities are considered to be related. For disambiguation purposes, keywords are added to the queries when obtaining the hit counts.

Matsuo et al (2004) apply the same method to find connections between members of a closed community of researchers. They gather person names from conference attendance lists to create the nodes of the network. The affiliations of each person are added to the queries as a crude form of named entity disambiguation. When a connection is found, the relation is labeled by applying minimal rules, based on the occurrence of manually selected keywords, to the contents of websites where both entities are mentioned.

A more elaborate approach to network mining is taken by Mika (2005) in his presentation of the *Flink* system. In addition to Web co-occurrence counts of person names, the system uses data mined from other—highly structured—sources such as email headers, publication archives and so-called Friend-Of-A-Friend (FOAF) profiles. Co-occurrence counts of a name and different interests taken from a predefined set are used to determine a person's expertise and to enrich their profile. These profiles are then used to resolve named entity co-reference and to find new connections.

Elson et al (2010) use quoted speech attribution to reconstruct the social networks of the characters in a novel. Though this work is most related regarding the type of data used, their method can be considered complementary to ours: where they relate entities based on their conversational interaction without further analysis of the content, we try to find connections based solely on the words that occur in the text.

Efforts in more general relation extraction from text have focused on finding recurring patterns and transforming them into triples (RDF). Relation types and labels are then deduced from the most common patterns (Ravichandran and Hovy, 2002; Culotta et al, 2006). These approaches work well for the induction and verification of straightforwardly verbalized factoids, but they are too restricted to capture the multitude of aspects that surround human interaction; a case in point is the kind of relationship between two persons, which people can usually infer from the text, but is rarely explicitly described in a single triple.

2.2 Sentiment Analysis

Sentiment analysis is concerned with locating and classifying the subjective information contained in a source. Subjectivity is inherently dependent on human interpretation and emotion. A machine can be taught to mimic these aspects, given enough examples, but the interaction of the two is what makes humans able to understand, for instance, that a sarcastic comment is not meant to be taken literally.

Although the general distinction between negative and positive is intuitive for humans to make, subjectivity and sentiment are very much domain and context dependent. Depending on the domain and context, a single sentence can have opposite meanings (Pang and Lee, 2008).

Many of the approaches to automatically solving tasks like these involve using lists of positively and negatively polarized words or phrases to calculate the overall sentiment of a clause, sentence or document (Pang et al, 2002). As shown by Kim and Hovy (2006), the order of the words potentially influences the interpretation of a text. Pang et al (2002) also found that the simple presence of a word is more important than the number of times it appears.

Word sense disambiguation can be a useful tool in determining polarity. Turney (2002) proposed a simple, but seemingly effective way to determine polarity at the word level. He calculates the difference between the mutual information gain of a phrase and the word ’excellent’ and of the same phrase and the word ’poor’.

3 Method, Data, and Annotation

3.1 Method

In contrast to most previous work regarding social network extraction, we do not possess any explicit record of the network we are after. Although the documents we work with are available online, the number of hyperlinks between them is minimal and all personal relations are expressed only in running text. We aim to train a system able to extract these relations and classify their polarity automatically using as little information as possible that is not explicitly included in the text, thus keeping the reliance on external resources as limited as possible.

We take the same approach with regards to the sentiment analysis part of the task: no predefined lists are supplied to the system and no word sense disambiguation is performed.

We take a supervised machine learning approach to solving the problem, by training support vector machines on a limited number of preclassified examples. We chose to use SVMs as a baseline method that has been shown to be effective in text categorization tasks (Joachims, 1998). We compare performance between joint learning, using one multi-class classifier, and a pipeline, using a single class classifier to judge whether an instance describes a relation, and a second classifier to classify the relations according to their polarity.

3.2 Data

We use the Biographical Dictionary of Socialism and the Workers’ Movement in the Netherlands (BWSA) as input for our system.[1] This digital resource consists of 574 biographical articles, in Dutch, relating to the most notable actors within the domain. The texts are accompanied by a database that holds such metadata as a person’s full name and known aliases, dates of birth and death, and a short description of the role they played within the Workers’ Movement. The articles were written by over 200 different authors, thus the use of vocabulary varies greatly across the texts. The length of the biographies also varies: the shortest text has 308 tokens, the longest has 7,188 tokens. The mean length is 1,546 tokens with a standard deviation of 784.

A biography can be seen as a summary of the most important events in a person’s life. Therefore, this type of data suits our purpose well: any person that the main character was closely related to, can be expected to appear in his or her biography.

In training our relation extraction system we look only at the relation from A to B and its associated polarity. The assumption that we make here is that by processing the BWSA in its entirety, making each of the 574 main characters person A once and harvesting all of their relations, we will get a full view of the existing relations, including the relation from B to A if A and B have a relation and B also has a biography in the BWSA.

We create one data set focused on a particular person who is prevalent throughout the data, namely Ferdinand Domela Nieuwenhuis (FDN). He started his career as a Lutheran priest, but lost his faith and pursued a career in socialist politics. After a series of disappointments, however, he turned to anarchism and eventually withdrew himself from the political stage completely, though his ideas continued to inspire others. We expect that the turmoil of his life will be reflected in his social network and the variety of relationships surrounding him.

As a first step in recreating Domela Nieuwenhuis’ network, we extract all sentences from the BWSA that mention the name ’Domela’, by which he is generally known. We exclude Domela’s own biography from the search. All but one of the ex-

[1] http://www.iisg.nl/bwsa/

tracted sentences, 447 in total, actually refer to Ferdinand Domela Nieuwenhuis. This sentence is removed, resulting in a total of 446 sentences spread over 153 biographies. Each sentence with a mention is expanded with additional context, to capture more clues than the sentence with the mention might hold. Preliminary tests showed that two sentences of context before the mention, and two sentences of context after the mention is sufficient. Often there is an introduction before a person is mentioned, and an elaboration on the relation after the mention. Figure 1 shows an example fragment.

However, since Domela was a rather controversial and a-typical figure, his network might not be a good representation of the actual relations in the data. Therefore, we create a second data set by randomly extracting another 534 sentences with their surrounding context from the BWSA that contain a named entity which is not the main entity of the biography. We aim to test which data set leads to better performance in finding and classifying relations across the entire community.

3.3 Annotation

All fragments in the Domela set were annotated by two human annotators, native speakers of Dutch, but unfamiliar with the domain of social history. They were asked to judge whether the fragment does in fact describe a relation between the two entities and, if so, whether the polarity of the relation from A to B is negative, neutral, or positive; i.e. whether person A has a negative, neutral or positive attitude towards person B.

With regards to the existence of a relation, the annotators reached an agreement of 74.9%. For the negative, neutral and positive classes they agreed on 60.8%, 24.2%, and 66.5%, respectively. All disagreements were resolved in discussion. The class distribution over the three polarities after resolution is shown in Table 1.

The generic set was annotated by only one of the annotators. The class distribution of this set is also shown in Table 1. It is roughly the same as the distribution for the A to B polarities from the Domela set.

Class	**Generic set**		**FDN set**	
	No.	%	No.	%
negative	86	16.1	74	16.6
neutral	134	25.1	87	19.5
positive	238	44.6	215	48.2
not related	76	14.2	70	15.7
total	534	100	446	100

Table 1: Class distribution

4 Relation Extraction and Classification

We train our system using LibSVM (Chang and Lin, 2001), an implementation of support vector machines. In training, the cost factor is set to 0.01 with a polynomial kernel type.

4.1 Preprocessing

First, all fragments and biographies are lemmatized and POS-tagged using Frog, a morpho-syntactic analyzer for Dutch (Van den Bosch et al, 2007). In a next step, Named Entity Recognition is performed with a classifier-based sequence processing tool trained on biographical data.

To identify the person to which a named entity refers, the name is split up into chunks representing first name, initials, infix and surname. These chunks, as far as they are included in the string, are then matched against the BWSA database. If no match is found, the name is added to the database as a new person. For now, however, we treat the network as a closed community by only extracting those fragments in which person B is one that already has a biography in the BWSA. At a later stage, biographies of people from outside the BWSA can be gathered and used to determine their position within the network.

4.2 Features

Co-occurrence counts: We calculate an initial measure of the relatedness of A to B using a method that is similar to Kautz et al (1997). The main difference is that we do not get our co-occurrence counts only from the Web, but also from the data itself. Since the domain of the data is so specific, Web counts do not accurately represent the actual distribution of people in the data. More famous people are likely to receive more attention on the Web than less famous people.

Ansing^{PER-A} and Domela Nieuwenhuis^{PER-B} were in written contact with each other since August 1878.
Domela Nieuwenhuis probably wrote uplifting words in his letter to Ansing, which was not preserved, after reading Pekelharing's report of the program convention of the ANWV in *Vragen des Tijds*, which was all but flattering for Ansing.
In this letter, Domela also offered his services to Ansing and his friends.
Domela Nieuwenhuis used this opportunity to ask Ansing several questions about the conditions of the workers, the same that he had already asked in a letter to the ANWV in 1877, which had been left unanswered.
Ansing answered the questions extensively.

Figure 1: English translation of an example fragment from the FDN set.

This is illustrated by Figure 2, where the number of times each person's name is mentioned within the BWSA is compared to the number of times he or she is mentioned on the Web.

We collect all possible combinations of each person's first names, initials and surnames (some are known by multiple surnames) and their aliases from the database and get the number of hits, i.e. the number of articles or webpages that contain the name, by querying the BWSA and Yahoo!. For each we derive 6 scores:

- *A-B*: the maximum hit count of all combinations of A ∩ B divided by the maximum hit count of A;
- *A-B(25)*: the maximum hit count of all combinations of A ∩ B within 25 words divided by the maximum hit count of A;
- *B-A*: the maximum hit count of all combinations of A ∩ B divided by the maximum hit count of B;
- *B-A(25)*: the maximum hit count of all combinations of A ∩ B within 25 words divided by the maximum hit count of B;
- *AB*: the maximum hit count of all combinations of A ∩ B divided by the maximum hit count of A plus the maximum hit count of B;
- *AB(25)*: the maximum hit count of all combinations of A ∩ B within 25 words divided by the maximum hit count of A plus the maximum hit count of B.

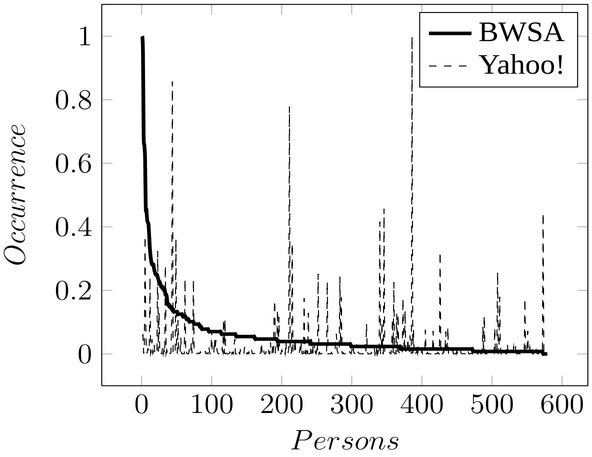

Figure 2: Fraction of maximum occurrence count for all 574 persons in the BWSA and on Yahoo!.

Set mention count: As an indication of the relatedness more specific to the text fragment under consideration, we add the number of times A or B is mentioned in the 5-sentence-context of the fragment, and the number of sentences in which both A and B are mentioned to the feature vector.

Lexical features: Preliminary tests revealed that keeping lemmatized verbs and nouns provided the best results, with mildly positive effects for prepositions and person names. All tokens outside these categories were not incorporated in the feature vector.

Person names are further processed in two ways: all mentions of person A and person B are replaced with labels 'PER-A' and 'PER-B'; all names of other persons mentioned in the fragment are replaced with label 'PER-X', where X is either the next available letter in the alphabet (anonymous) or the person's unique ID in the database (identified).

We create four variants of both the generic data set and the FDN data set: one that represents only verbs and nouns (VN), one that also includes prepositions (VNPr), one that includes anonymous person names (VNP-a) and a last one that includes identified person names (VNP-i). Each set is split into a training and a test set of respectively 90% and 10% of the total size. We test our system both with binary features and with tf.idf weighted features.

5 Results and Evaluation

5.1 Binary versus Tf.idf

Figure 3 shows the 10-fold cross-validation accuracy scores on the joint learning task for each of the training vector sets using binary and tf.idf weighted features. We take the majority class of the training set as our baseline. In all cases we observe that unweighted binary features outperform weighted features. These results are in line with the findings of Pang et al (2002), who found that the occurrence of a word is more important than its frequency in determining the sentiment of a text.

Regarding the different feature sets, the addition of prepositions or person names, either anonymous or identified, does not have a significant effect on the results. Only for the VNP-a set the score is raised from 47.86 % to 48.53 % by the inclusion of anonymous person names.

5.2 Co-occurrence

We perform a second experiment to assess the influence of adding any of the co-occurrence measures to the feature vectors. Figure 4 displays the results for the VN set on its own and with inclusion of the set mention counts (M), the BWSA co-occurrence scores (B) and the Yahoo! co-occurrence scores (Y).

For the generic set, we observe in all cases that the co-occurrence measures have a negative effect on the overall score. For the FDN set this is not always the case. The set mention counts slightly improve the score, though this is not significant. The remainder of the experiments is performed on the vectors without any co-occurrence scores.

5.3 Joint Learning versus Pipeline

Table 2 lists the accuracy scores on the training sets on both the joint learning task and the pipeline. Only for the FDN set does the system perform better on the two-step task than on the single task. In fact, the FDN set reaches an accuracy of 53.08 % in the two-step task, which is 6.55 % higher than the majority class baseline and the highest score so far.

The system consistently performs better on the joint learning task for the generic set. Further investigation into why the pipeline does not do well on the generic set reveals that in the first step of the task, where instances are classified on whether they describe a relation or not, all instances always get classified as 'related'. This immediately results in an error rate of approximately 15%. In the second step, when classifying relations into negative, neutral or positive, we observe that in most cases the system again resorts to majority class voting and thus does not exceed the baseline.

Even for the FDN set, where the pipeline does outperform the joint learning task, the difference in accuracy between both tasks is minor (0.22-0.96 %). We conclude that it is preferable to approach our classification problem as a single, rather than a two-step task. If the system already resorts to majority class voting in the first step, every occurrence of a name in a biography will be flagged as a relation, which is detrimental to the precision of the system.

5.4 Generic versus FDN

Although the classifiers trained on both sets do not perform particularly well, the FDN set provides a greater gain in accuracy over the baseline. The same is shown when we train the system on the training sets for both data sets and test them on the held out test sets. For the generic set, the VNP-a feature set provides the best results. It results in an accuracy of 50% on the test set, with a baseline of 48.2%.

For the FDN data set, none of the different feature sets performs better than the others on the joint learning task. In testing, however, the VNP-a set proves to be most successful. It results in an accuracy of 66.7%, which is a gain of 4.5% over the baseline of 62.2%.

To test how well each of the sets generalizes over the entire community, we test both sets on each

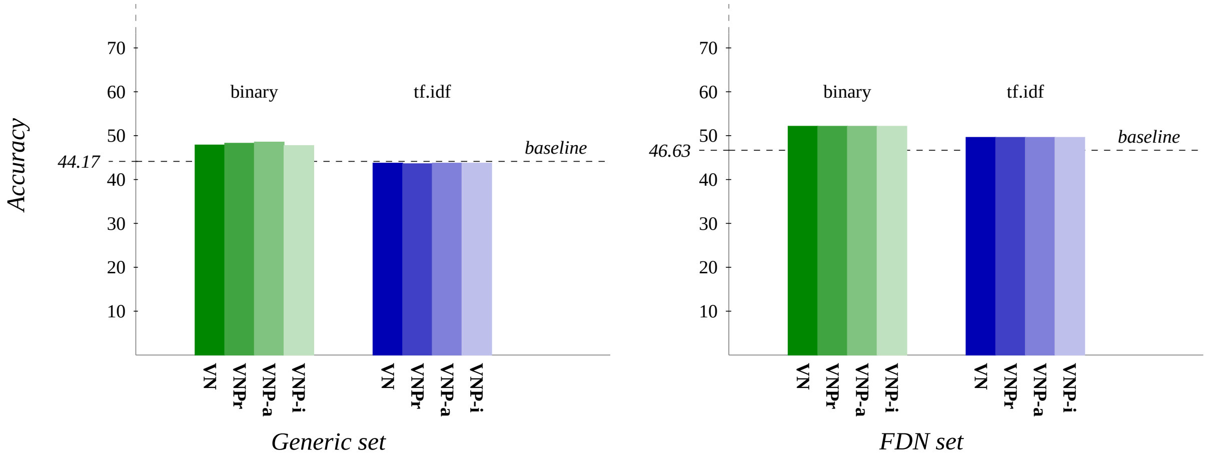

Figure 3: Binary versus weighted features.

	Generic set			**FDN set**		
	joint	*pipeline*	*baseline*	*joint*	*pipeline*	*baseline*
VN	47.92	45.83	44.17	52.12	52.83	46.63
VNPr	48.33	46.88	44.17	52.12	**53.08**	46.63
VNP-a	**48.54**	46.88	44.17	52.12	52.34	46.63
VNP-i	47.71	45.83	44.17	52.12	52.59	46.63

Table 2: Accuracy scores on training sets (10-fold cross-validation) for both the joint learning task and the pipeline.

other. Training on the generic set and testing on the FDN set results in an accuracy of 45.3% with a baseline of 48.2%. Doing the same experiment vice versa results in an accuracy of 44.8% with a baseline of 44.6%. Examining the output reveals that both systems resort to selecting the majority class (’positive’) in most cases. The system that was trained on the FDN set correctly selects the ’negative’ class in a few cases, but never classifies a fragment as ’neutral’ or ’not related’. The distribution of classes in the output of the generic system shows a bit more variety: 0.2% is classified as ’negative’, 10.1% is classified as ’neutral’ and 89.7% is classified as ’positive’. None of the fragments are classified as ’not related’. A possible explanation for this is the fact that the ’not related’ fragments in the FDN set specifically describe situations where the main entity is not related to Ferdinand Domela Nieuwenhuis; these fragments could still describe a relation from the main entity to another person mentioned in the fragment and therefore be miss-classified.

5.5 Evaluation

To evaluate our system, we process the entire BWSA, extracting from each biography all fragments that mention a person from any of the other biographies. We train the system on the best performing feature set of the generic data set, VNP-a. In order to filter out some of the errors, we remove all relations of which only one instance is found in the BWSA.

The resulting network is evaluated qualitatively by a domain expert on a sample of the network. For this we extracted the top-five friends and foes for five persons. Both rankings are based on the frequency of the relation in the system’s output. The lists of friends are judged to be mostly correct. This is probably due to the fact that the positive relation is the majority class, to which the classifiers easily revert.

The generated lists of foes are more controversial. Some of the lists contain names which are also included in the list of friends. Of course, this is not

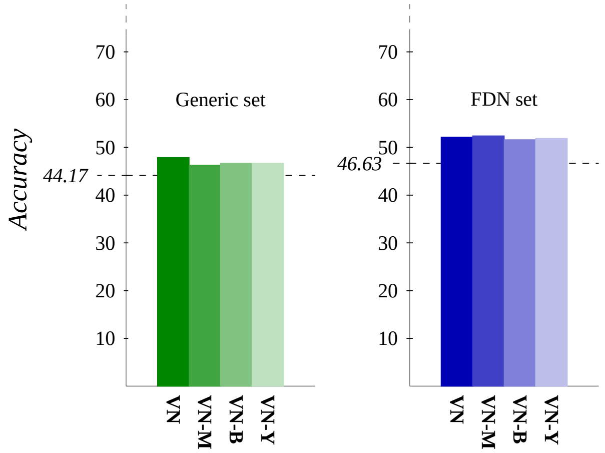

Figure 4: Comparison of co-occurrence features: M = set mention counts, B = BWSA co-occurrence, Y = Yahoo! co-occurrence.

necessarily a sign of bad system performance: we do not count time as a factor in this experiment and relationships are subject to change. 25% of the listed foes are judged to be completely wrong by the expert judge. 10% are not so much enemies of the main entity, but did have known political disagreements with them. The remaining 65% are considered to be plausible as foes, though the expert would not have placed them in the top five.

6 Discussion and Future Research

Our case study has demonstrated that relations between persons can be identified and labeled by their polarity at an above-baseline level, though the improvements are minor. Yet, the utility of the classifications is visible in the higher-level task of constructing a complete social network from all the classified pairwise relations. After filtering out relations with only one attestation, a qualitative analysis by a domain expert on frequency-ranked top-five lists of friends and foes yielded mostly correct results on the majority class, ’positive’, and approximately 65% correct on the harder ’negative’ class. If we would not have used the classifier and guessed only the majority ’positive’ class, we would not have been able to build ranked lists of foes.

In discussions with domain experts, several extensions to our current annotation scheme have been proposed, some of which may be learnable to some usable extent (i.e. leading to qualitatively good labelings in the overall social network) with machine learning tools given sufficient annotated material. First, we plan to include more elaborate annotations by domain experts that discriminate between types of relationships, such as between family members, co-workers, or friends. Second, relationships are obviously not static throughout time; their polarity and type can change, and they have a beginning and an end.

We aim at working with other machine learning methods in future expansions of our experimental matrix, including the use of rule learning methods because of their interpretable output. Another direction of research, related to the idea of the improved annotation levels, is the identification of sub-networks in the total social network. Arguably, certain sub-networks identify ideologically like-minded people, and may correspond to what eventually developed into organizations such as workers unions or political organizations. When we are able to link automatically detected temporal expressions to initializations, changes, and endings of relationships, we may be able to have enough ingredients for the automatic identification of large-scale events such as the emergence of a political movement.

References

Antal van den Bosch, Bertjan Busser, Sander Canisius and Walter Daelemans. 2007. *An efficient memory-based morphosyntactic tagger and parser for Dutch*. Selected Papers of the 17th Computational Linguistics in the Netherlands Meeting, Leuven, Belgium, 99–114.

Chih-Chung Chang and Chih-Jen Lin. 2001. *LIBSVM: a library for support vector machines*. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Aron Culotta, Andrew McCallum and Jonathan Betz. 2006. *Integrating probabilistic extraction models and data mining to discover relations and patterns in text*. Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL) 2006, 296–303.

Walter Daelemans, Jakub Zavrel, Ko van der Sloot and Antal van den Bosch. 2010. *TiMBL: Tilburg Memory*

Based Learner, version 6.3, Reference Guide. ILK Research Group Technical Report Series no. 10-01.

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 2010, 138–147.

Thorsten Joachims. 1998. *Text categorization with support vector machines: Learning with many relevant features*. Proceedings of ECML-98, 10th European Conference on Machine Learning 1998, 137-142.

Henry Kautz, Bart Selman and Mehul Shah. 1997. *The hidden web*. AI Magazine, volume 18, number 2, 27–36.

Soo-Min Kim and Eduard Hovy. 2006. *Automatic identification of pro and con reasons in online reviews*. Proceedings of the COLING/ACL Main Conference Poster Sessions, 483–490.

Yutaka Matsuo, Hironori Tomobe, Koiti Hasida and Mitsuru Ishizuka. 2004. *Finding social network for trust calculation*. European Conference on Artificial Intelligence - ECAI 2004.

Peter Mika. 2005. *Flink: Semantic web technology for the extraction and analysis of social networks*. Web Semantics: Science, Services and Agents on the World Wide Web, volume 3, number 2-3, 211–223.

Bo Pang, Lillian Lee and Shivakumar Vaithyanathan. 2002. *Thumbs up? Sentiment classification using machine learning techniques*. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 79–86.

Bo Pang and Lillian Lee. 2008. *Opinion mining and sentiment analysis*. Foundations and Trends in Information Retrieval, vol. 2, number 1-2, 1–135.

Deepak Ravichandran and Eduard Hovy. 2002. *Learning Surface Text Patterns for a Question Answering System*. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL) 2002.

Peter D. Turney. 2002. *Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews*. Proceedings of the Association for Computational Linguistics (ACL), 417-424.