UPPSALA
UNIVERSITET

Methods for lean, precision-oriented, and targeted coreference resolution

Daniel Lindmark

Uppsala University
Department of Linguistics and Philology
Språkteknologiprogrammet
(Language Technology Programme)
Bachelor's Thesis in Language Technology

July 16, 2012

Supervisors:
Joakim Nivre, Uppsala universitet
Fredrik Olsson, Gavagai AB

Abstract

This thesis presents the work of evaluating and improving an existing module for lean and targeted coreference resolution in a system for multilingual and dynamic, social media text analytics. In order to evaluate the current method a mimicking module was developed and an evaluation corpus in three languages was compiled including three public persons as targets. With focus on maintaining precision while increasing recall, the evaluation revealed some weak and strong points within the system concerning the use of certain categories of strings, referred to as expansions, assumed to be coreferential with a target and used for retrieval of sentences. In particular, the system is most vulnerable when epithets for the target are used. Two methods for improving the system were proposed and evaluated. The first makes use of simple sentence distances in order to filter out sentences containing non-coreferential expansions and in the second a named entity recognizer is applied for the same purpose. Results showed that both methods and especially a combination of them had a positive effect on the overall result but that the first is only beneficial for certain kinds of expansion types while the implementation of the second is impractical given the demands of speed and multilinguality on the system.

Contents

Acknowledgements

I'm thankful to Joakim Nivre for support and guidance regarding the structure and content of this thesis. I would also like to thank everyone at Gavagai for your welcoming attitude, for giving me the opportunity to experience Language Technology "IRL", and for the nice coffee! In particular I would like to thank Fredrik Olsson for being an excellent supervisor.

1 Introduction

"Big data" is a recent buzzword within Information Technology and refers to the huge amounts of data presently facing the industry, that are big not only in terms of bytes but also in variety and velocity. Some sources point out that 90% of the data in the world today has been created in the last two years alone.[1] A large part of this data is due to the so called "social data revolution" brought about by social media networks, blogs, microblogs, forums and wikis. The textual part of this data alone, provides great opportunities as well as challenges for large scale information extraction. The opportunities spring from the amounts of data available to be analyzed, spotting trends and opinions in real time. Many challenges arise from the fact that the majority of this data is unstructured, heterogeneous and from many different languages. Many times this data is created or changed so fast that in order to be available for searching and analytics upon creation or update, it cannot be handled by traditional methods and software. Other challenges, always present when handling natural language with the help of computers, are to handle the inherent ambiguity of human language and to deal with concepts and references that require real world knowledge. Such concepts and references would typically regard entitites such as persons, companies and products.

Within the field of Natural Language Processing (NLP), the process of identifying references in human language to real world entities is referred to as *reference resolution*. The vast research on the subject clearly shows its importance and difficulty. Nilsson (2010) motivates this research mainly because resolving referential expressions can serve as a useful preprocessing step for many applications. Specifically, such applications may include systems for Information Extraction (IE), Passage Retrieval, Question Answering (QA), Automatic Summarization and Abstracting, Machine Translation (MT) and Dialogue Systems. Considering the work presented in this thesis, the applications would include systems for Opinion Mining and Text Analytics.

The difficulties of computerized reference resolution notably arise from the fact that human real world knowledge needed for the task is hard to model. Relying on lexical knowledge only is not sufficient. Accordingly, in practically all NLP approaches to coreference resolution some kind of pre-processing tool for tagging and/or parsing, more or less extensively, is used on the data to analyze. Still, NLP approaches differ a lot as to what extent linguistic, semantic and world knowledge are used as well as to what methods are applied to utilize this information. However, most research has considered English as the language of interest and as in much other NLP research, data used for development and evaluation are often very limited in size and genre.

[1] http://www-01.ibm.com/software/data/bigdata/

In contrast to this, this thesis concerns coreference resolution in a multilingual and dynamic, social media data environment. This is a major challenge considering today's consistent flow of big data and will inevitably put constraints on applicable pre-processing methods as well as on the choice of algorithms used to adress the task. Thus, earlier research strategies may not be adequate. First of all, since large amounts of streaming data has to be analyzed, the use of time consuming NLP tools for pre-processing is practically impossible. Secondly, considering languages that are typologically very different from English e.g. Chinese and Arabic, it is unlikely that algorithms developed for English coreference resolution will yield the same success rates for them. Similarly, social media text is an emerging "genre" that has not yet been extensively treated in NLP research. What we might expect though, is that streaming social media data normally is not as structured and coherent as the corpus data used in earlier research.

The work of this thesis is carried out in association with Gavagai AB, a company that develops *Ethersource*, a system built for dynamic real-time text analytics. *Ethersource* is used for the tracking and monitoring of signals that comprise attitudes and sentiments, to be found in large streams of textual and symbolic data, towards known targets. On a typical day, between 2.5 and 5 millions of social media posts in multiple languages pass through the system.

1.1 Purpose

The purpose of this thesis is twofold. First of all, it is to evaluate a simple coreference resolution model that is already implemented in *Ethersource* and point out the pitfalls and costs associated with this method. Secondly, the purpose is to find possible improvements to this model, where an improvement means to increase recall of utterances referring to a target, without sacrificing precision.

When it comes to improvements of the model, the demands on *Ethersource* to handle huge amounts of streaming data in a vast number of different languages puts a number of constraints on the choice of methods to apply. These constraints are:

- The approach should be scalable and therefore lean in terms of time efficiency and memory footprint.
- The approach should be applicable to all languages currently available in *Ethersource* (Arabic, Greek, English, Finnish, French, Hindi, Russian, Swedish, Ukranian and Chinese).

In addition to these constraints, as noted above but in need to be emphasized, in order to be an improvement, the approach must not decrease precision.

1.2 Outline

The outline of this thesis is as follows. Chapter 2 gives a background to the work in this thesis including terminology of essential concepts, history and background of the field as well as presenting the coreference resolution challenge

in *Ethersource* in more detail. Chapter 3 describes the work done in order to evaluate the current coreference resolution model of the system and presents the result of this work. Chapter 4 continues with proposing methods for improvements to the model and evaluating. Chapter 5 contains a discussion of the findings in both chapter 3 and 4. A final conclusion is given in chapter 6.

2 Background

This chapter begins by introducing some basic but important terminology concerning the task of coreference resolution in linguistics and NLP. The first section is followed by a brief overview of the history and state of the art of coreference resolution as a field within NLP. For a more thorough exposé, see for example Mitkov (2002) and Ng (2010). The chapter ends with a more detailed description of the previously introduced system under scrutiny in this work, *Ethersource*. In particular, the current algorithm of the coreference module referred to as "coref light" is outlined and the method explained. A note on related work is given at the end.

2.1 Coreference and anaphora resolution

2.1.1 Linguistic terms and concepts

Reference is the symbolic relationship that a linguistic expression has with a real world entity. Particular linguistic expressions are said to *refer* to certain objects. C*oreference* exists when multiple expressions in a discourse are used to refer to the same real world entity. For instance, in example 1 the expressions "Romney", "a former Massachusetts governor", "who", "himself", and "the GOP's inevitable nominee" are all said to be referring expressions and thus *coreferential* with the *referent*, in this case the US republican candidate Mitt Romney.

(1) The victories will provide an important boost for Romney, a former Massachusetts governor who has sought to cast himself as the GOP's inevitable nominee.

Jurafsky and Martin (2009) define *reference resolution* as "the task of determining what entities are referred to by which linguistic expression." Ng (2008) further specifies *coreference resolution* as "the problem of identifying which *mentions* (i.e., noun phrases) refer to which real-world *entities*." However, the coreference resolution task is not always limited to "objects referred to by noun phrases or pronouns [. . .]. But coreference involving events, expressed via verbs or nominalised verb forms, is also common" (Humphreys et al., 1997). Typically though, coreference resolution signifies the identification of *coreference chains* i.e. classes of expressions realized by definite noun phrases (NPs) and pronouns, that refer to the same entity.

Anaphora resolution then denotes the more specific process of resolving references to previously introduced entities in a discourse. It can thus be seen as a sub-task of coreference resolution although it has been at the center of the

field from the beginning. Instead of focusing on the whole chain of referring expressions, the aim in anaphora resolution is to distinguish the *antecedent* of a particular *anaphoric*[1] expression pointing back to it. Naturally, there is also the possibility in discourse to use a referring expression that precedes the entity it refers to. In this case the referring expression is said to be *cataphoric*. However, *cataphora* is a much less frequent phenomenon than anaphora. Commonly, two different kinds of anaphora are distinguished, namely *nominal anaphora* and *pronominal anaphora*. The resolution of the former concerns finding the antecedent of a definite NP or a name, while resolution of the latter denotes the specific case of finding the earlier introduced antecedent of a pronoun. In addition to these two types of anaphora there exist some less commonly treated varieties, among which *verb anaphora* and *zero anaphora* may be the least uncommon. An example of verb anaphora is shown in example 2 where the single verb *did* points back to the anteceding verb phrase *stop running*.

(2) As the father told the child to *stop running* it *did* so.[2] (Nobre, 2011)

Zero anaphora is a peculiar phenomenon that occurs when an anaphor seems to be hidden or invisible since no overt representation exists. According to Mitkov (2002), the most common sub type is *zero pronominal anaphora* in which case "the anaphoric pronoun is omitted but is nevertheless understood". Kılıçaslan et al. (2009) presents figures for the ratio of zero pronouns in Turkish texts ranging from 62 – 75% of all pronouns in a text. Mitkov (2002) further lists Spanish, Italian, Portuguese, Polish, Chinese, Japanese, Korean and Thai as languages where zero pronominal anaphora is prevalent, and goes on stating that "NLP applications covering these languages cannot circumvent the problem". An example of zero pronominal anaphora is shown in the Spanish sentence in example 3 where the symbol Ø shows the position of the omitted pronoun.

(3) Ø Se llama Fernando Vargas. Ø Es un indígena de etnia guaraní. ([He] is called Fernando Vargas. [He] is a native of Guaraní descent).

2.1.2 History and state of the art

Since coreference resolution and anaphora resolution are so closely related, it is not easy to draw a line between their development. Both disciplines emerged as fields of study within NLP as far back as in the 1960s (Ng, 2010; Mitkov, 1999) and have influenced each other ever since, and in much recent research they are more or less presented as a joint problem. As pointed out by Clark and Gonzáles-Brenes (2008), while being a well studied subject, results particularly in coreference resolution have been disappointing until recently. Following the development of the field, it is seen that many early approaches relied heavily on rules or discourse theories, thus being referred to as *discourse based* or *knowledge intensive* making use of extensive linguistic and domain knowledge. Moving on to the 90s, the trend shifted to *knowledge poor* methods, where

[1] From the ancient Greek word ἀναφορά: "a carrying back in an upward direction".

[2] It may be argued that the last word "so" should be part of the anaphoric expression although this is not considered in the work from which the example is derived. However, this would rather be an example of adverb anaphora.

usage of linguistic and domain knowledge is limited (Küçük, 2005). Yet, most of the early approaches can be termed *knowledge based*, since they all somehow make use of linguistically motivated, handcrafted constraints and rules. Modern approaches have embraced more data driven methods motivated by the progress in machine learning, statistical methods and clustering techniques. Another notable difference between early research and modern, pointed out by Mitkov (2001), is that most earlier algorithms depended on manual pre-editing, often were simulated and applied to perfectly syntactic input. As a consequence, comparisons between evaluations of early and modern methods are not always reliable.

2.1.3 Knowledge based approaches

Centering theory

Many of the early, discourse based approaches belong to the family of *centering theory* (Grosz et al., 1983, 1995). These theories build upon the idea that within an utterance (typically a sentence or a clause), there are entities that are central, or topically prominent, and therefore defined as *centers*. In order to find and rank the centers in an utterance, a syntactic parse, together with information about grammatical functions, is needed. A set of rules, selectional restrictions and constraints govern the choice of entities marked as centers. For instance, entities in subject position are preferred over those in object position. Adjacent utterances are then examined in a sequence, looking for continuation, retaining or shifting of centers (i.e. topic). Centering theory at its core, comprises a model of discourse coherence. An implementation of one of the variants of centering theory, the Left-Right Centering Algorithm (Tetreault, 1999) reported 72.4% correctly resolved pronouns.

Hobbs' "naive" approach

One of the most influential of the knowledge poor approaches, already presented in the late 70s, is Hobbs algorithm, or Hobbs' "naive" approach (Hobbs, 1978). Hobbs presented a simple and efficient algorithm for resolution of pronominal anaphora that still serves as a benchmark and a baseline in much research. In addition to a syntactic parser, the algorithm requires a gender and number checker. In short, the resolution procedure is a breadth-first, left-to-right search where a parse tree is traversed relatively to a certain pronoun, looking for NPs that are candidate antecedents to this pronoun. Constraints on the way the search is performed as well as constraints for agreement of gender, person and number, along with additional semantic constraints, are used in order to decide if the candidate is a valid antecedent. Hobbs presented an impressive success rate of 91.7% for the algorithm, when all constraints were used. However, it should be noted that Hobbs never implemented the algorithm, the procedure was only manually simulated. Moreover, data was constructed manually resulting in perfectly syntactic input. Later implementations of the algorithm, using automatic preprocessing methods have shown a decrease of performance down to 71% or slightly lower (Nilsson, 2010).

Mitkov's algorithm

Another influential and more recent approach to anaphora resolution is Mitkov's robust knowledge poor algorithm (Mitkov, 1998), later implemented in the fully automatic system MARS (Mitkov et al., 2002). The original algorithm was developed with the explicit intention of being practically useful for real NLP problems and therefore refrained from using parsing and complex, semantic or syntactic analysis. It was designed to rely on post-edited output from a POS-tagger and an NP-extractor in order to find and score possible antecedents of pronouns in a text, based on several indicators. Examples of positive indicators of anaphoricity for candidate antecedents are:

- First NP in a sentence
- NP immediately following an indicating verb (from a pre-defined set)
- Lexical reiteration of a NP (string matched or NPs with the same head)
- NP in section heading
- Collocation pattern match: *press the key, press it*

Examples of negative indicators of anaphoricity for candidate antecedents are:

- Indefinite NP
- Prepositional NP

The algorithm works by first limiting search space to the two preceding sentences for a pronoun and then selecting candidates that fulfill gender and number constraints. Candidates are then scored by adding together the indicators that apply and the one with the highest score is selected as antecedent. Mitkov (2002) reports a success rate of 89.7% for the original algorithm which was evaluated on technical manuals.

In MARS, the fully automatic implementation of the algorithm, no post editing is done, instead a parser is used. In addition, more constraints are applied and new indicators added together with a number of other improvements. Yet, the MARS system reported much lower but more realistic success rates for a working system than the original.

Mitkov's algorithm has proven to be successful for a number of other languages than English, such as Polish, Arabic, Bulgarian, French and Portuguese. This success is partly explained by the fact that some of the indicators, e.g. distance, repetition and collocation patterns, are language and genre independent (Nilsson, 2010). The highest success rates are found for languages with more diverse, yet restricted agreement rules, and when the algorithm is language specifically modified.

2.1.4 Data driven approaches

MUC and ACE

The Message Understanding Conferences (MUC) in the late 90s, laid the foundation for the rise of many data driven approaches with their focus on creating useful annotated data sets for a number of different tasks. The conferences were organized as competitions where these tasks were adressed in order to find new,

or improve, methods for Information Extraction (IE). In MUC-6 and MUC-7, the tasks of named entity recognition (NER) and coreference resolution had been introduced and resulted in annotated evaluation corpora, guidelines for annotation of named entitites and coreferential relations, as well as evaluation metrics for both tasks. Even if no standard metric exists for coreference resolution evaluation, metrics from these conferences usually serve as a point of reference together with other metrics. This is also the case for the data itself, the MUC-7 corpus still being among the ten most purchased LDC[3] corpora. MUC-6 introduced the notion of, and the convention of calling a referring expression, a *markable*. In the resulting corpus, every such markable is annotated, together with pairwise coreference or anaphoric relations between markables, using SGML.

Another important and more recent initiative for the development of resources for IE is the Automatic Content Extraction (ACE)[4] program that started in 2000, and which over the years has resulted in a number of annotated data sets for training and evaluation, primarily in English, but also in Chinese and Arabic. The differences between MUC and ACE are not only that the latter defines *mentions* instead of *markables* but also that these can be defined as one of seven types. Furthermore, not only entitites were considered for annotation but also events and relations.

Supervised and unsupervised approaches

Nilsson (2010) divides the data driven field into two branches, unsupervised and supervised. A supervised approach typically relies on feature vectors representing a mention itself and relations between mention pairs. The feature vectors are then fed to a machine learning algorithm in order to automatically classify mentions and detect coreference chains. In a typical unsupervised approach, coreference resolution is considered a clustering task. Thus, based on similarity between feature vectors describing each NP, these are partitioned into clusters supposed to be the equivalent of a coreference chain. Unsupervised approaches are attractive since they do not require annotated data.

A good example of an influential, supervised approach is Soon et al. (2001), which was "the first machine learning based system to offer performance comparable to that of nonlearning approaches." In this approach, a small training data set is used in which coreference chains of NPs have been annotated. A "pipeline" of NLP modules are used for extensive preprocessing and marking up with rich lexical, morphological, syntactic and semantic information. Then, feature vectors are constructed based on 12 features found for a markable pair *i* and *j*:

1. DIST – sentence distance between *i* and *j*
2. I_PRONOUN – *i* is a pronoun (t/f)
3. J_PRONOUN – *j* is a pronoun (t/f)
4. STR_MATCH – string match between *i* and *j* (t/f)
5. DEF_NP – *j* is a definite noun phrase (t/f)

[3]The Linguistic Data Consortium: http://www.ldc.upenn.edu/
[4]http://projects.ldc.upenn.edu/ace/

6. DEM_NP – *j* is a demonstrative noun phrase (t/f)
7. NUMBER – number agreement between *i* and *j* (t/f)
8. SEMCLASS – *i* and *j* belong to the same semantic class (t/f or unknown)
9. GENDER - *i* and *j* have the same gender (t/f or unknown)
10. PROPER_NAME - both are proper names (t/f)
11. ALIAS – if one is an alias of the other (t/f)
12. APPOSITIVE – *j* is in apposition to *i* (t/f)

After feature vector generation, both positive and negative examples of coreferential chains are used as input to a machine learning algorithm that learn from the feature vectors. In order to apply the classifier to unseen data, the new data has to first go through the pipeline of pre-processing.

The state of the art

Many different machine learning techniques and algorithms have been extensively and successfully employed in recent years for coreference resolution and the supervised approaches still show the best performance. A current trend in coreference resolution is to experiment with unsupervised approaches. These have generally not reached the same performance as fully supervised, but are not far behind. For comparison, Ng (2008) presents F-scores, using the MUC-scoring algorithm, for an unsupervised system that are on average 3.9% lower compared to a fully supervised model and in Lang et al. (2009) this gap is only 2.57%. In fact, Raghunathan et al. (2010) presents an unsupervised method outperforming a state-of-the-art supervised system on one data set with an F-score of 80.8% compared to 80.4% using the B^3 metric (Bagga and Baldwin, 1998).

2.2 Ethersource

Ethersource is a text analytics technology developed for real time, dynamic text analytics. Its purpose is to compute and track relations between terms and symbols in streaming language data, typically originating from social media such as Facebook, Twitter, Flickr, blogs, microblogs and internet forums.

Generally, *Ethersource* is based on a vector space model. In such a model, words are represented as high dimensional vectors based on their distribution within a document or a set of documents. The relative directions of the context vectors of words are assumed to indicate semantic similarity (Sahlgren, 2005). This means that if we find words that, by some similarity measure, are close to each other in a vector space, they will be words that appear in the same contexts, such as synonyms and collocations etc. More specifically, *Ethersource* is built upon a random indexing (RI) model, where context vectors of words or documents are *accumulated*, incrementally based on contextual occurrence. RI is said to be "an efficient, scalable and incremental alternative to standard word space methods" (Sahlgren, 2005).

Apart from efficiency and scalability, the biggest advantage of *Ethersource* in comparison to many of its competitors, using statistical or knowledge based models, is the possibility to handle multilinguality, language change and variety,

without having to retool the knowledge base and update the model. Currently, the system is able to perform sentiment and attitude analysis in Arabic, Greek, English, Finnish, French, Hindi, Russian, Swedish, Ukranian and Chinese. With more data available in more than 50 languages, and with the intention to handle more of the available languages, as well as to increase data volumes, all the inherent advantages of *Ethersource* will be crucial.

The current system relies on pre-defined *targets*. These targets, referring to entities such as persons, companies or even currency pairs, are concepts, realized by a set of strings that are selected by domain experts and used for the retrieval of relevant documents. The chosen documents are further examined in a process called *polarization* where documents are split into utterances and those containing the target are analyzed with respect to a pre-defined *pole* representing attitudes or sentiments.

Imagine that we are interested in analyzing attitudes of violence within social media towards the current Swedish Prime Minister. The target would then typically be represented by the full name *Fredrik Reinfeldt* and spelling variants thereof and might also include epithets deemed by domain experts as very reliable descriptions, e.g. *Statsminister Reinfeldt (Prime Minister Reinfeldt)*. All these strings of full names and descriptions are assumed to be references to the target, and their context, i.e. the sentence they are found in, as eligible for polarization. Next, the pole would stand for a concept of violence, and every utterance containing the target is scored according to the degree of violence it comprises. This score is estimated by the occurence of expressions that *Ethersource* define as violent.

2.2.1 The coreference challenge

A notable shortcoming with using narrowly defined descriptions of a target is that the limited set of target terms, while ensuring a certain measure of precision, limits the recall of possible other utterances referring to the target. In short, we miss out on valuable information that maybe could have given us a better overall signal of sentiments and attitudes. This is the essence of the coreference resolution challenge in *Ethersource*. The solution is not as straightforward as to simply include every possible string of reference to the target within the set of target strings since that would result in retrieval of a huge number of documents, and hence utterances, that would be completely irrelevant. Clearly, including the 47th most common Swedish first name[5] *Fredrik* as a target term would over-generate, not to mention the stupidity of including the pronoun *han*(he). Even using his rather uncommon last name *Reinfeldt* in the same way would not cause the same amount of damage but we would still end up with documents regarding any member of the Reinfeldt family, e.g. the politician and (ex?)wife of the Prime Minister, Filippa Reinfeldt, or the high-media profile son, Erik Reinfeldt.

A less naive approach, which is the current approach of the system, is not to expand the set of target descriptions and retrieve more documents, but rather to work on retrieving more utterances from the fairly reliable documents, always including at least one mention of a target description. This is accomplished by

[5] http://spraakbanken.gu.se/statistik/lbfnamn.phtml

defining a set of *expansion* terms, i.e. first names, last names and possibly epithets, that are included in the search space once the target matching documents have been retrived, but before polarization. Still, pronouns are banned even in the expansion set and the approach is by no means bullet proof and there is still the risk of over-generation and loss of precision.

2.2.2 Current algorithm - "Coref light"

The following very straightforward algorithm for coreference resolution is implemented in the current system. Note that polarization is included as a last step in the algorithm but that it is not part of the actual process of resolving coreferences.

Algorithm 1 Current coreference resolution algorithm in *Ethersource*

Given a set of documents, a target and a set of expansions

for all documents **do**
  if document contains target **then**
    extract all sentences containing target or expansion
  else
    skip document
  end if
end for

Polarize extracted sentences

2.2.3 Related work

Since much research in the area is focused on classical coreference resolution it is hard to find approaches similar to that of this thesis. Large scale projects to handle "more than a million mentions" exist but they still use a lot of preprocessing and seem to handle speed and scalability issues with the help of hardware or techniques that can make use of parallellism (Singh et al., 2011). Optimization issues concerning hardware, programming or advanced mathematical models is lamentably not within the scope of this thesis but may well be an appropriate topic in possible further studies.

There seem to exist no test beds of practical use in this project. Equally, because of the difference from traditional coreference resolution, markup strategies and evaluation metrics, origination from the MUC conferences or other coreference annotated corpora, do not apply. At least not any more than that of marking up if a mention is coreferential or not.

3 Evaluation

This chapter describes the work done in order to evaluate the current coreference module in *Ethersource* and in order to investigate the pros and cons of using an expansion set in the way as described above and in the algorithm. First, the method used for the evaluation is introduced followed by a description of its implementation and evaluation metrics. Finally the results are presented in the form of three case studies for three different targets in three different languages.

3.1 Method

In order to evaluate the current algorithm a few steps were needed. First of all, a strategy was set for assessing the usefulness of different expansions. Secondly, a program was developed and used for both collecting data and to evaluate the current method, as well as for doing experiments in order to improve methods. The program was entirely built in Java which is the preferred developing language for *Ethersource*. The program can be divided into three parts:

1. A module with the ability to mimic the behaviour of the current "coref light" module[1]. The "mimicking" in this case refers only to the process of coreference resolution and does not include polarization. Thus, the evaluation only regards the success of the coreference resolution part of the algorithm and does not evaluate the effects this may have on the overall sentiment analysis.

2. A method for corpus compilation. A corpus of blog posts had to be compiled, containing annotated data, in order to measure the performance of the current method. The data to work with originates from blog posts in three different languages, Swedish, English and Spanish, in order to somehow address multilinguality, at least on a small scale.

3. A framework for evaluation.

3.1.1 Division of expansion terms

As seen, the current system makes use of expansion terms in order to retrive more sentences than just the ones that contain a certain target string (normally

[1]For a number of reasons it was more practical to develop a separate evaluation environment than to evaluate directly on *Ethersource*.

the full name of the target). As noted previously, these are typically single instances of the first or last name of the target. Other expansion terms that could be called "epithets" are also used in conjunction with the names but are not as easily defined. The epithets can be very different from target to target but an epithet is normally a defining term that is used frequently in media or on the web to describe the target or disambiguate it from other named entities.

For evaluation of this approach, in order to better estimate the effect of different expansion terms, these were divided into four categories for easier handling, namely:

1. *lastname*
2. *firstname*
3. *epithet*
4. *pronoun*

This division will provide a better view of how each category contribute to the overall recall and precision. Note that pronouns are now included in the expansion set even if they are not used as expansions in the current system. Now they are included in order to measure the gain or damage of including them.

3.1.2 Description of the mimicking module

The following outline describes the mimicking procedure of the program given a set of retrieved documents in one of the three languages, already known to contain target strings:

1. Initially, the set of target terms and expansion terms for a certain expansion category are loaded in the form of regular expression strings.

2. All documents are loaded into memory, retrieving the document content as well as information about the source i.e. the URL to the website it originates from.

3. For each document, tokenization and sentence splitting is done. In the original system, the Apache Lucene[2] software library is used for this purpose but here this is accomplished through operations on input strings by regular expressions. It is believed that the result of this "home built" solution will not differ very much from the original approach since the output is very similar.

4. Document content and source information are stored in a data structure together with the index of the document according to it's order of appearance in the stream. This data structure also stores information about targets and expansion terms. The search for target and expansion strings is also made by pattern matching using the `java.util.regex` library.

5. Finally, all sentences containing expansions are printed out together with their indexes and meta information in XML formatted markup.

[2]http://lucene.apache.org/

3.1.3 Corpus compilation

The data used to build the evaluation corpus is derived from freshly extracted social media data supplied by one of Gavagai's data providers. Three targets were selected corresponding to the three languages Swedish, English and Spanish. Selected targets are the Swedish Prime Minister *Fredrik Reinfeldt*, the US republican candidate and senator *Ron Paul* and the Venezuelan President *Hugo Chávez*. Considering expansion categories *lastname* and *firstname* for these three targets it would not be too hard to see that these would include "reinfeldt", "paul", "chávez" and "fredrik", "ron" and "hugo." Examples of epithets for the targets would be, for Fredrik Reinfeldt: "partiledare" (party leader), for Ron Paul: "congressman", and for Hugo Chávez: "comandante" (commander). Naturally, not all possible pronouns in the three languages are included in the *pronoun* category, only those that apply i.e. those that agree in gender or number with the target. As will be seen in the Spanish case, since not all languages share the same constraints and rules when it comes to gender and number agreement, including pronouns can be a risky business. All strings that are used for the three targets in the evaluation corpus are listed in appendix A.

A standard unix/linux `grep` command was initially used to extract the documents matching the allowed target strings and the three data sets were also chosen from matching time spans but differ substantially in size as seen in table 3.1.

Table 3.1: Extracted and annotated data for each target

Target	size of data (approx.)	# documents	# annotated documents
Fredrik Reinfeldt	940 KB	387	276
Ron Paul	9.5 MB	2 917	262
Hugo Chávez	24.5 MB	5 808	227

Not all of the data is used in the case of *Paul* and *Chávez*, partly because of the time consuming task of markup, partly because of the desire to use comparable amounts of data. The smallest data set reaches up to a document index of 385 and the two other data sets were matched approximately to that figure. However, documents containing less than two sentences are discarded and the highest index is not necessarily the same as the number of annotated documents. The numbers of annotated documents for each target are shown in the rightmost column in table 3.1.

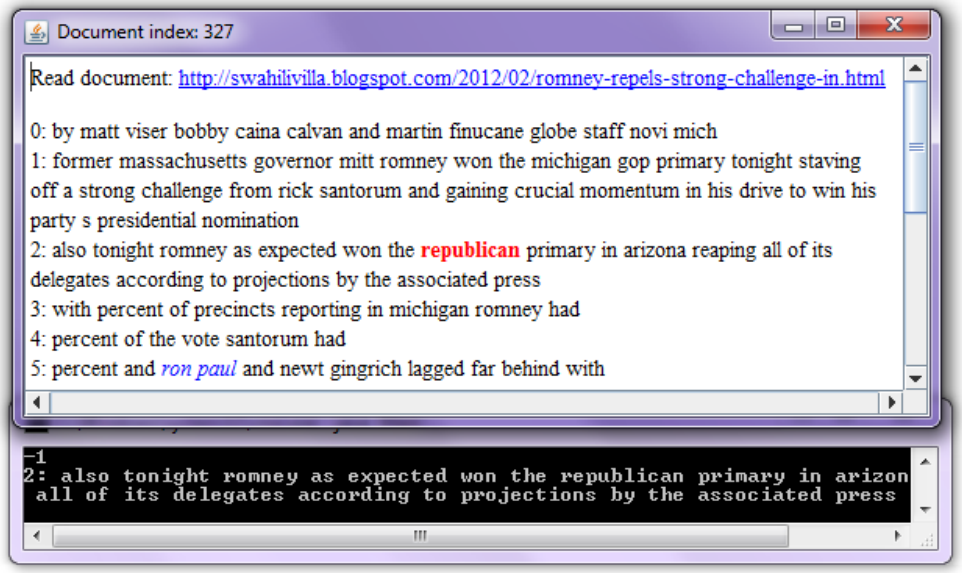
Annotation

There is an annotation method integrated in the program in order to construct the evaluation corpus described in table 3.1 and in greater detail in table 3.2. With the help of this method, annotation of sentences containing expansion terms but not containing any target term was done for all documents within a chosen index span.

The annotation procedure is as follows:

1. In a loop going through a desired index span of documents, each document content is presented in its entirety in a pop-up window. For the purpose, a

Figure 3.1: Annotation with the help of a `JTextPane`



`JTextPane`, from the `javax.swing` library is used, with the text marked up with HTML code for easy view of targets and expansion as well as the URL for the document. An example is shown in figure 3.1.

2. For every sentence in the document where an expansion is found, an input of either 1 or -1 is given through the command line depending on whether the expansion is coreferent or not with the target. Documents containing less than two sentences are discarded.

3. Finally, information is merged and the sentences together with their indexes and meta information are printed out in XML formatted markup.

Since the expansions were divided into four categories, this procedure was repeated for each category, resulting in four different XML files for each language with a variation in the number of sentences in each as specified in table 3.2. The result of the corpus annotation is a collection of sentences that are marked up as either positive (1) or negative (-1). Since annotation is done on the sentence level, coreference relations will be presented rather coarse grained. For example, if there are multiple occurrences of one type of expansion within one sentence, only the first (i.e. leftmost)[3] will be regarded. Thus, if one or more of the other expansions refer to a different entity than that of the first, this information will be lost and the result possibly somewhat distorted. The reason for this is mainly a matter of practicality and lowering the annotation load.

Another issue to point out is that there is some redundancy within the corpus since the same information may appear on multiple sites, on blogs or forums. Yet, this is the nature of the web and it can be argued that the data only models reality.

[3]It surely would have been more natural to mark a sentence as positive if at least one of the expansions were coreferential. However, this was considered after annotation was already done.

Table 3.2: # annotated sentences for each expansion

Target	*lastname*	*firstname*	*epithet*	*pronoun*
Fredrik Reinfeldt	302	51	81	519
Ron Paul	286	18	403	1 288
Hugo Chávez	583	13	571	1 041

3.1.4 The evaluation framework

The evaluation framework consists of a class with methods to aggregate and compare data from the XML formatted output files containing indexed sentences with a certain type of expansion. Initially, all of the annotated sentences are aggregated and each unique index is counted in order to find the intersection of all positive as well as negative sentences. This information will then serve as a key to the total number of positive sentences. The result from single files and aggregates of files with different expansions can then be compared to this file by matching the indexes in order to assess the impact of single expansions and combinations. When experimenting on the model, the resulting output is compared in the same way to this key.

Precision and recall

The metrics used for evaluating the success of the current algorithm on the annotated data are precision and recall, where the desire is to retrieve as many positive sentences as possible without lowering precision.

In this setting, by recall is meant the ratio of retrieved sentences deemed as positive divided by the total number of positive sentences.

$$\text{recall} = \frac{\text{\# retrieved positive sentences}}{\text{\# positive sentences}}$$

Precision then is the ratio of retrieved sentences deemed as positive divided by the total number of retrieved sentences, either positive or negative.

$$\text{precision} = \frac{\text{\# retrieved positive sentences}}{\text{\# retrieved sentences}}$$

As mentioned above, the total number of positive sentences is given by the intersection of all positive sentences found in the annotated corpus.

3.2 Results

This section presents an evaluation of the current coreference resolution algorithm in the form of three case studies for the three chosen targets introduced above. Results are given for each target in terms of precision and recall of retrieved positive sentences. In the upper part of each table listing the figures, the impact for each expansion is shown. Below the single categories are also

presented figures for combinations of *lastname* with each of the other expansion terms. As seen, it seems motivated to always include *lastname* concerning both precision and recall. Note that when two expansions are found in the same sentence, it is possible that one is positive and the other is negative. In this case, the choice has been to count the sentence as a positive one which affects precision positively. Below the combinations, figures regarding the current system are given, referred to as "current system", in which the categories *lastname*, *firstname* and *epithet* are included. Note that no sentences including a target term is taken into account in order to focus on the effects of using expansions. At the bottom of each table the results for combining all expansions, including the *pronoun* category are presented.

3.2.1 Case study 1 - Fredrik Reinfeldt, Swedish

Figures for precision and recall concerning the target *Fredrik Reinfeldt* are shown in table 3.3. As might be expected, using category *lastname* gives the best precision. The inclusion of *firstname* with *lastname* may seem a little but not highly motivated since the gain in recall is only 6% while some precision is lost. After all, "Fredrik" is a common name. The use of epithets show an ever larger drop in precision. This uncovers the dangers of using epithets. There is always the possibility to use more general terms that will capture too much. In this case 92% of the sentences including the expression *partilerare* (party leader), are false positives. Regarding pronouns, clearly, avoiding them seems highly motivated.

Table 3.3: Fredrik Reinfeldt - Impact on precision and recall for single categories of expansions and combinations

Category	*P*	*R*
lastname	**93%**	57%
firstname	82%	9%
epithet	55%	9%
pronoun	36%	38%
lastname + firstname	91%	63%
lastname + epithet	86%	65%
lastname + pronoun	55%	**89%**
Current system	85%	70%
All expansions	56%	100%

The current system seems to produce a good trade-off between precision and recall in the Swedish case. However, if precision would be found to be more important for the polarization process it would perhaps be enough to only use *lastname*.

3.2.2 Case study 2 - Ron Paul, English

Moving on to another language then, I let one of the bloggers introduce the somehow more "complicated" target:

"Ron Paul is not a real name. A real name consists of a first name and a last name, not of two first names."

Although I may disagree that "Ron Paul" is not a valid name, it seems that we are bound for trouble since the string representing *lastname* in this case can also be a quite common English first name. Really, this only makes the results for the chosen target all the more interesting to investigate. Figures for precision and recall concerning the target *Ron Paul* are shown in table 3.4. The drop in precision compared to *Fredrik Reinfeldt* may be explained by this double nature of the *lastname* and since *Ron* is a common name too, the effect is similar. The use of single first names however, does not seem very common in comparison to last names and full names.

Table 3.4: Ron Paul - Impact on precision and recall for single categories of expansions and combinations

Category	*P*	*R*
lastname	**86%**	64%
firstname	35%	2%
epithet	7%	9%
pronoun	12%	42%
lastname + firstname	83%	66%
lastname + epithet	38%	70%
lastname + pronoun	23%	**94%**
Current system	37%	72%
All expansions	20%	100%

A look at the epithets then reveals that the term "republican" is "the bad guy". It matches too much since it can not only a stand as a noun but also as an adjective. In the majority of the wrong sentences, this is the case. The current method seems to perform poorly if terms like these and tricky names are used.

3.2.3 Case study 3 - Hugo Chávez, Spanish

In Spanish, we encounter the zero pronoun as well as pronouns attached to verbs. In our case these "peculiarities" would perhaps mean that there would be fewer instances of pronouns and thus not lead to as much increase in recall as for the other languages. Also, the fact that the pronoun *su* is not bound to a specific gender or number but can mean both *his*, *her* or *their* may cause some problems. In general, the use of pronouns as expansions for Spanish seems tricky since there are many interfering forms and they are used in many different ways. For that reason, the masculine direct object pronoun *lo* was excluded from the expansion set. Figures for precision and recall concerning the target *Hugo Chávez* are shown in table 3.5.

Even in this case there is a significant drop in precision when using epithets. The epithet producing most errors is *venezolano* (venezuelan) which again, is an example of an expression that is often used adjectivally and thus overgenerates.

Table 3.5: Hugo Chávez - Impact on precision and recall for single categories of expansions and combinations

Category	*P*	*R*
lastname	**99%**	58%
firstname	46%	0.6%
epithet	57%	33%
pronoun	35%	40%
lastname + firstname	98%	59%
lastname + epithet	80%	82%
lastname + pronoun	55%	84%
Current system	80%	**83%**
All expansions	55%	100%

The highest precision again is shown to be given by *lastname*, which this time is next to perfect. *Chávez* seems to have made a strong last name for himself while the first name returns very little recall.

Surprisingly, for the *pronoun* category both precision and recall are in line with the Swedish case which may signify that the limited set of pronouns chosen are adequate.

4 Improvements

This chapter adresses the second purpose of the thesis, to find ways to improve the results presented in the evaluation above. Two different approaches will be tested and evaluated with regard to their usefulness for the coreference resolution task as well as their practicality concerning the demands of the system.

4.1 Method

As previously stated, in practically all research on coreference resolution, syntactic, morphological or semantic parsing together with NER are seen as a prerequisite in order to apply an algorithm, learn a classifier or perform clustering. In contrast to this, the demand of the present system and the difference of this work compared to "traditional" coreference resolution is that of avoiding the use of expensive NLP tools. The question is then more of finding extremely simple features that have proven to be distinguishing in determining coreference relations between entities and referring terms.

One such simple feature is the sentence distance of a certain expression from the antecedent and is often used as one of many features for coreference resolution. According to Jurafsky and Martin (2009), sentence distances, modeling recency, have proven to be of good use in the case of pronoun resolution. The first method to experiment with will therefore regard sentence distances and the overall idea is that the further away an expansion is from a preceding target it is less recent and therefore less likely to be coreferential.

This simple approach will probably be proven too simple on its own and a second approach is proposed which will unfortunately compromise the earlier mentioned constraints of speed and practicality. NER has proven to be of great importance for coreference resolution and therefore, opting for a "light-weight" markup procedure where we make use of a NER tagger for partial "parsing" may not seem far fetched. It would at the same time shed some light over what is lost in performance by not using such tools as well as the cost in terms of speed and memory of using one.

How then, would NER benefit coreference resolution in the present setting? Basically, the idea is to identify person names other than those referring to the target, assuming that this would make it possible to identify negative sentences. This way of applying NER is a bit different from traditional NER since we are only interested in the specific task of person name recognition. The assumption is based partly on observations during annotation, that person names other than that of the target often appear close to expansions that are non-coreferential.

These names can be seen as "false" targets and therefore to be marked as belonging to a new category called *falsetarget*.

Both the first approach, using distances and the second approach, using NER can be seen as filters that are used to get rid of unwanted sentences. In order to visualize the effects of these filters in a good way, figures of what is left after the filters have been applied will be shown, as either remaining "true positives" or remaining "false positives". By true positives are meant those sentences that are positive and false positives thus refers to the remaining negative sentences containing non-coreferential expansions. These measures are related to those of precision and recall. A higher percentage of remaining true positives correlates with a higher recall and a lower percentage of remaining false positives correlates with a higher precision.

After being treated separately, the results of combining both filters will be evaluated according to measures of precision and recall and in comparison with the current approach and scalable alternatives.

4.2 Distances

In order to model recency, sentence distances between targets and expansions were computed and considered. The tables below list the results for all three targets. The figures show the percentage of the remaining coreferential sentences (true positives), and the remaining negative sentences (false positives), after removing sentences containing expansions that are further away than two sentences from a preceding target.

4.2.1 Results

Table 4.1: Fredrik Reinfeldt - remaining true positives (TP) and false positives (FP) when discarding expansions further away than two sentences from a preceding target

Category	TP	FP
lastname	20%	36%
firstname	19%	0%
epithet	31%	22%
pronoun	57%	7%

Results were not overwhelmingly convincing for all types of expansions but interesting regarding some and figures for the English data in table 4.2 and the Spanish in 4.3 both point to a similar pattern to the Swedish data in table 4.1. Overall, for category *lastname*, the distance from a target seems to have no correlation with coreference. Concerning category *firstname*, conclusions are hard to make because of the sparseness of data. However, for categories *epithet* and *pronoun*, distances do seem to matter. Regarding them, true coreferential mentions most often occur in the same sentence as the target or in the two closest following sentences.

At least regarding the English and Swedish data sets, category *pronoun* seems to gain the highest precision by applying the distance filter. In the case of the

Table 4.2: Ron Paul - remaining true positives (TP) and false positives (FP) when discarding expansions further away than two sentences from a preceding target

Category	TP	FP
lastname	32%	30%
firstname	66%	18%
epithet	70%	10%
pronoun	38%	6%

Spanish results in 4.3, a possible explanation to the lower loss of false positives compared to the *epithet* category might be the inclusion of the possessive pronoun *su*, which is not bound either to gender or number.

Table 4.3: Hugo Chávez - remaining true positives (TP) and false positives (FP) when discarding expansions further away than two sentences from a preceding target

Category	TP	FP
lastname	20%	33%
firstname	0%	0%
epithet	33%	10%
pronoun	22%	12%

4.3 Using NER

The benefits of using named entity recognition (NER) together with coreference resolution has been pointed out before. Mitkov (2002, p. 38) stresses that "the detection of NP anaphors requires at least partial parsing in the form of NP extraction. A named entity recognizer, and in particular a program for identifying proper names, could be of great help at this stage".

In order to perform person name recognition and retrieve a list of *falsetarget* terms the *Apache OpenNLP*[1] toolkit is used.This toolkit implements machine learning techniques and is developed for addressing a number of different NLP tasks. For the specific purpose of person name recognition, the `NameFinderME` class is used. It requires a trained model for every language to be handled. Thus, three different models, Swedish[2], English and Spanish[3], were used in this project. The Spanish model is trained on *CoNLL02* shared task data. The origin of training data for the other two models is not specified. The OpenNLP name recognizer (or the model) is not 100% accurate, as supposedly no name recognizer is. For instance, during tests, the expression "América Latina" was found to be considered a person name instead of a geographic location.

With the help of *OpenNLP*, person name recognition is performed as a pre-processing step on the entire amount of data in each set, not only on the annotated part. This makes it easier to estimate the real cost in speed of the

[1]http://opennlp.apache.org/

[2]Thanks to Svetoslav Marinov for providing the Swedish model.

[3]http://opennlp.sourceforge.net/models-1.5/

process. Before any person name recognition is done all strings matching target terms are filtered out. A few issues had to be adressed regarding `NameFinderME`. First of all, the method extracting the names generates token N-grams of varying sizes, from one token up to several. These tokens can sometimes be even singleton letters. In order to prevent overmatching, N-gram sizes were limited to two tokens, only considering full names and not just single last names and first names or long names with three tokens or more. Secondly, especially with the two larger data sets, a full retrieval of all possible full names, even with only two tokens, resulted in large numbers of instances. A search for all of these slows down the program considerably. Also, using all instances proved to overgenerate. To adress this, a threshold for the frecuency of occuring names was set, depending on the size of the data.

OpenNLP was chosen because of its availability and ease of use. Another NER system, free or commercial, may do the job better and faster. Nevertheless, in order to somehow estimate the practicality of using a tool like this, figures of speed of performing NER within the evaluation system will be presented. As noted above, another important issue to adress concerning NER and speed is the effect on the speed of the sentence extraction process when including a search for all the identified *falsetarget* strings. All tests were performed on a fairly modern home computer with an Intel Core 2 CPU processor with a speed of 1.87 GHz and with 4GB RAM installed. The operating system is Microsoft Windows 7 (64 bit).

4.3.1 Results

Results show that using NER for the identification of negative sentences can often improve precision while not losing extensively in recall. At the same time there seem to be some differences between expansion categories within a data set and between categories in the three data sets. The tables under each target show the percentage of the remaining positive sentences (true positives), and the remaining negative sentences (false positives), after removing sentences containing a *falsetarget*, i.e. a full name other than that of the target.

Fredrik Reinfeldt

For the Swedish data with a size of 879 KB after removing target strings, NER took around 23 seconds to run in order to collect a list of *falsetarget* terms for 387 documents. The speed per document is thus around 0.06s. A threshold of at least two mentions was used resulting in a list of 213 full names. Sentence extraction with retrieval of targets, expansions and interfering names takes 18 seconds. A sentence extraction without retrieval of *falsetarget* strings only takes 3 seconds.

For the Swedish set, the overall result seems to improve and with better NER precision it would have been even better. For category *lastname* the unwanted sentences referring to the wife "Filippa Reinfeldt" of the target can now be identified and removed, although at the same time those referring to both "Fredrik" and "Filippa" are filtered out as well, lowering recall. Problematic sentences for *firstname* are those containing only single names i.e. "Fredrik". There seems to be no simple straightforward way to further filter these out.

Table 4.4: Fredrik Reinfeldt - Remaining true positives(TP) and false positives(FP) when discarding sentences containg a *falsetarget*

Category	TP	FP
lastname	79%	4%
firstname	65%	33%
epithet	87%	33%
pronoun	91%	79%

For the *epithet* category, damage caused by the single term "partiledare" (party leader) is now limited, yet half of the remaining false sentences contain that expansion term. The others are difficult ones containing single first or last names or indefinite use of epithets. For pronouns, distances to *falsetarget* do not prove to be as useful as distances to targets and the result is much more scattered.

Ron Paul

The time taken for NER to find names for the whole English set of 9.2MB after removing target strings is 5 minutes and 43 seconds for 2 917 documents and substantially slower than for the Swedish data. This is understandable since the data size is around ten times bigger. Yet, the time per document to perform NER has literally doubled and is now 0.12s. Using all retrieved names for the search of *falsetarget* terms makes sentence extraction take around 10 minutes so therefore I use a threshold of 5 for the occurence of a name which brings the figure down to 3 minutes and 23 seconds and also yields better results. A sentence extraction without *falsetarget* strings takes 28 seconds.

Table 4.5: Ron Paul - Remaining true positives(TP) and false positives(FP) when discarding sentences containg a *falsetarget*

Category	TP	FP
lastname	89%	57%
firstname	93%	21%
epithet	89%	57%
pronoun	96%	84%

For *lastname* the results seem not as convincing as for the Swedish case in removing false sentences, but with better NER precision, some more instances could have been correctly identified, the wife of Ron Paul for example, referred to as *Mrs Paul*. For the *firstname* category results are rather pleasing but single names are still a challenge. For *epithet* results are not as impressive due to the damage caused by one *epithet* term in particular namely *republican*. Better NER could improve results somewhat but removing *republican* from the expansion set would be better.

Hugo Chávez

The time taken for NER to handle the Spanish data set of 23.9MB after removing target strings is 15 minutes and 17 seconds for 5 808 documents.

The speed per document is now 0.16s. A threshold of 5 for the occurence of a name is used here as well. Sentence extraction with retrieval of all terms take 4 minutes and 16 seconds. A sentence extraction without retrieval of *falsetarget* strings takes 30 seconds.

Table 4.6: Hugo Chávez - Remaining true positives(TP) and false positives(FP) when discarding sentences containg a *falsetarget*

Category	TP	FP
lastname	83%	0%
firstname	76%	84%
epithet	90%	70%
pronoun	87%	87%

For Spanish, results do not look as convincing except for *lastname*, but again, in almost all cases, at least for *firstname* and *epithet*, there are cases of missed person names. For *epithet*, the term *venezolano* is still responsible for 40% of the unwanted sentences. On the other hand, *venezolano* stands for around 20% of the correct sentences, yet these are often retrieved using *firstname* or *lastname*.

4.4 Combined filters and scalable alternatives

Let us then return to the previously used metrics of precision and recall and see the results of combining the filters in comparison with the current approach of the system and that of using only the *lastname* category as well as in comparison with two other scalable approaches where only the filter of target distances is used.

Combining both filters

When mixing the different expansion sets and allowing both filters to be utilized, there were some configurations that overall seemed to perform better than others. The best result concerning *lastname*, seemed to be given when using only the *falsetarget* filter without the distance filter. This pruned category is called FT_ln in the table. The same applies to *firstname* and likewise this category is called FT_fn. For *epithet* and *pronoun* filters for both *falsetarget* and target distance were used and are thus referred to as FTTD_ep and FTTD_pr. Figures are not given for all separate expansions but only for two combinations, one without FTTD_pr and one including these filtered pronouns.

Scalable alternatives

The scalable alternatives to the combined approach are all those that do not apply the *falsetarget* filter for any expansion. That includes all configurations listed in section 3.2 but also all possible configurations that make use of the target distance filter. For practical purposes only the best configurations of the alternatives are presented. There are two configurations shown that make use of the target distance filter. One includes both the filtered *epithet* category referred

to as TD_ep and the filtered *pronoun* category referred to as TD_pr and the other only includes the TD_ep category.

Table 4.7: Fredrik Reinfeldt - Precision and recall with combined filters and scalable alternatives

Category	*P*	*R*
FT_ln + FT_fn + FTTD_ep	**98%**	52%
FT_ln + FT_fn + FTTD_ep + FTTD_pr	92%	72%
ln + fn + TD_ep	90%	65%
ln + fn + TD_ep + TD_pr	88%	**84%**
lastname	93%	57%
lastname + firstname	91%	63%
Current system	85%	70%

Table 4.8: Ron Paul - Precision and recall with combined filters and scalable alternatives

Category	*P*	*R*
FT_ln + FT_fn + FTTD_ep	79%	63%
FT_ln + FT_fn + FTTD_ep + FTTD_pr	69%	75%
ln + fn + TD_ep	73%	70%
ln + fn + TD_ep + TD_pr	66%	**83%**
lastname	**86%**	64%
lastname + firstname	83%	66%
Current system	37%	72%

Table 4.9: Hugo Chávez - Precision and recall with combined filters and scalable alternatives

Category	*P*	*R*
FT_ln + FT_fn + FTTD_ep	97%	58%
FT_ln + FT_fn + FTTD_ep + FTTD_pr	88%	62%
ln + fn + TD_ep	96%	68%
ln + fn + TD_ep + TD_pr	89%	72%
lastname	**99%**	58%
lastname + firstname	98%	59%
Current system	80%	**83%**

Results

The results in the tables show that the performance of the current approach may very well be improved by combining the two filters. It may even be improved only by using the distance filter. For two of the targets, the goal of increasing recall while maintaining precision is reached either by the target distance filter alone or by combining both filters. For the Spanish target, precision is affected

positively by using any of the approaches. The best precision is again attained when excluding pronouns, even when they are filtered. The results also show that the filtered approaches with the highest precision figures in two cases are lower than that of category *lastname* alone, while recall figures are on the same level as *lastname*.

5 Discussion

In this thesis an evaluation of the current coreference resolution algorithm has been presented in the form of three case studies, for three different public persons in three different parts of the world. Comparison between usage of person names and referring expressions in different languages can be daunting. Names have different patterns of distribution across the world, a first name or last name as well as their combination can be either in frequent use or not. Thus precision and recall can be affected for these reasons. Also, people may talk differently about certain types of persons and there may be cultural differences in the way people write and speak about public persons. There is also the question of the number and nature of epithets to use and it would be extremely hard to find epithets that are equivalent in order to compare their affect on recall and precision. With that in mind it is still believed that a comparison of three cases, since evidently it is not possible to investigate how all names or descriptions behave, will shed some light on what methods or expansion terms work or not, not really despite the differences but rather because of them.

The evaluation of the current algorithm shows that the use of the *lastname* category is by far the most useful category on its own if high precision is seen as most desirable. The results for *firstname* show, a little surprisingly perhaps, that its usage can be questioned since it does not affect recall very much but still lowers precision. Similarly, using *epithet* strings may give the same effect and in some cases lower precision dramatically. When it comes to the *epithet* category the crucial part seems to be deciding what strings to include in order not to overmatch since one certain string alone can lower precision significantly. A simple strategy to prevent this is to avoid chosing epithets that can be used as adjectives. For the data sets in the evaluation, this would have proven useful but it remains to be seen if this strategy is too simple since language is not as simple as it may seem looking only at a few examples.

At the same time, regardless of the effect of a certain *epithet* or a common name, the identification of interfering full names in a sentence seems to make sense. Using an automatic NER tool may thus improve precision but it may still miss out on some sentences and it may also lead to overmatching and even the filtering out of too much. As also seen, it may be beneficial to perform distance filtering for epithets. In fact, combining both filtering methods seems to be the best for the *epithet* category.

Pronouns seem more predictable in one way, yet unpredictable in many ways. In the experiments it has proven useful to limit the retrieval of sentences containing a pronoun to a distance of 1 or 2 sentences following a target, just as for the epithets. Although results in precision for pronouns are better after this it may still be desirable to restrict the inclusion of pronouns until more research

has been done and improvements found. With a more fine grained annotation of the corpus it may be easier to see certain patterns.

Both the proposed methods of improvement have seen to influence performance of coreference resolution positively, although a filter based on sentence distances alone does not impact precision as much as when using both filters. However, the filter based on target distances is a scalable method and in two cases it actually manages to accomplish the second purpose of this thesis, to improve the current method by increasing recall of utterances referring to a target, without lowering precision. In further work, it would be interesting to investigate the possibility to make use of paragraph identification in combination with distances, since such information has been helpful for coreference resolution in other settings.

Regarding the second filter, due to the restriction of the current system, it may not be a possible alternative because of the time it takes to first perform NER and then retrive the *falsetarget* sentences. The test results indicate speed issues when scaling. The memory demands to store all *falsetarget* terms may also be to costly. In addition, a person name recognizer may require a model for every language treated. If such models do not exist, they have to be trained on annotaded data which will demand time and additional work. Even then, it is impossible to expect a recognition of names with full precision.

If it could be done in an a more efficient way than with NER, one possible solution would perhaps be to concentrate on finding full name patterns that match either the *firstname* or the *lastname* together with a complementing name string. However, this approach will likely require some kind of "stop word" list to prevent overgeneration and such a list would in turn also require time and additional work in order to be maintained. Possible solutions for efficient name recognition or additional tagging might require hardware or software that make use of parallelism in order to speed up the process. Using the inherent word space model in *Ethersource* might be another alternative. A RI approach to NER has been suggested by Algotsson (2007). However, it may be argued that there are three most practical ideas at the moment:

1. If an overall improvement of the current method is enough, continue to use categories *lastname*, *firstname* and *epithet*, include category *pronoun* and apply a target distance filter for strings in the *epithet* and *pronoun* category.

2. In order to focus more on precision and when dealing with languages with not so strict pronoun agreement rules, do as above but exclude the *pronoun* category.

3. If precision is crucial, make use of the *lastname* category exclusively.

6 Conclusion

This project was embarked on with the object to evaluate the current coreference resolution algorithm in *Ethersource*, a system for text analytics on streaming social media data, and thereafter propose improvements to this method where an improvement is the same as increasing recall while maintaining precision.The demands on the system in terms of data volumes and multilinguality put constraints on these methods to be lean. Thus, the use of pre-processing tools normally applied for the task was seen as undesirable from the beginning because of speed and practicality issues and the character of the data being handled. Useful NLP tools for all languages treated by *Ethersource* simply do not exist.

In order to evaluate the current system a corpus with social media data in three languages was compiled, annotated with coreferential relations between a target and possible referring expansion strings. The corpus was created and used for evaluation with the help of a program built in Java, also developed to mimic the current module, as well as containing functionality to experiment with and evaluate the current approach to coreference resolution in *Ethersource*.

The evaluation showed that the current system is much more vulnerable when using the first name of a person and descriptions and epithets of the same in order to retrive more sentences than when only using the last name. The choice of epithets should be subject to restrictions and more research on what these restrictions should be is encouraged.

To find ways to improve the coreference module in *Ethersource*, two methods were tested and evaluated. Initially, a simple distance measure between target sentences and sentences with expansions was applied in order to look for patterns that could reveal coreferentiality. Such patterns were found to exist and useful to a certain point for filtering out unwanted sentences in the case of expansions such as pronouns and epithets. This method proved to be a scalable approach that in some cases actually fulfilled the aim of increasing recall while maintaining precision in comparison with the current method.

In order to adress the other types of expansions, last names and first names, the stepping away from the initial intention of abstaining from tagging the data was necessary. A "light-weight" procedure using NER was carried out using a free NLP tool, *OpenNLP*, with separate models for Swedish, English and Spanish. Using NER to identify sentences containing full names and filter them out proved generally to be of advantage for precision in relation to recall but may not be a feasible alternative for the current system given the constraints.

Bibliography

Gustav Algotsson. Automatic pronoun resolution for swedish. Master's thesis, the School of Computer Science and Engineering, Royal Institute of Technology, Stockholm, 2007.

Amit Bagga and Breck Baldwin. Algorithms for scoring coreference chains. In *Proceedings of the Linguistic Coreference Workshop at The First International Conference on Language Resources and Evaluation (LREC'98)*, pages 563–566, 1998.

Jonathan H. Clark and José P Gonzáles-Brenes. Coreference resolution: Current trends and future directions. *Language and Statistics II Literature Review*, page 14, 2008.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. Providing a unified account of definite noun phrases in discourse. In *Proceedings of the 21st annual meeting on Association for Computational Linguistics*, ACL '83, pages 44–50, Stroudsburg, PA, USA, 1983. Association for Computational Linguistics.

Barbara J. Grosz, Scott Weinstein, and Aravind K. Joshi. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21:203–225, 1995.

Jerry R. Hobbs. Resolving pronoun references. *Lingua 44*, pages 311–338, 1978.

Kevin Humphreys, Robert Gaizauskas, and Saliha Azzam. Event coreference for information extraction. In *Proceedings of a Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, ANARESOLUTION '97, pages 75–81, Stroudsburg, PA, USA, 1997. Association for Computational Linguistics.

Daniel Jurafsky and James H. Martin. *Speech and Language Processing (2nd Edition)*. Pearson Education International, Upper Saddle River, NJ, USA, 2009.

Yılmaz Kılıçaslan, Edip Serdar Güner, and Savaş Yıldırım. Learning-based pronoun resolution for turkish with a comparative evaluation. *Computer Speech & Language*, 23(3):311 – 331, 2009.

Dilek Küçük. A knowledge-poor pronoun resolution system for turkish. Master's thesis, The Graduate School Of Natural And Applied Sciences, Middle East Technical University, 2005.

Jun Lang, Bing Qin, Ting Liu, and Sheng Li. Unsupervised Coreference Resolution with HyperGraph Partitioning. *Computer and Information Science*, 2(4): 55–63, November 2009.

Ruslan Mitkov. Robust pronoun resolution with limited knowledge. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 2*, ACL '98, pages 869–875, Stroudsburg, PA, USA, 1998. Association for Computational Linguistics.

Ruslan Mitkov. Anaphora resolution : The state of the art. Technical report, University of Wolverhampton, Wolverhampton, 1999.

Ruslan Mitkov. Outstanding issues in anaphora resolution (invited talk). In *Proceedings of the Second International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing '01, pages 110–125, London, UK, 2001. Springer-Verlag.

Ruslan. Mitkov. *Anaphora resolution*. Studies in Language and Linguistics. Longman, 2002.

Ruslan Mitkov, Richard Evans, and Constantin Orăsan. A new, fully automatic version of mitkov's knowledge-poor pronoun resolution method. In *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, Mexico City, Mexico, February, 17 – 23 2002.

Vincent Ng. Unsupervised models for coreference resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 640–649, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

Vincent Ng. Supervised noun phrase coreference research: the first fifteen years. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 1396–1411, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

Kristina Nilsson. *Hybrid Methods for Coreference Resolution in Swedish*. PhD thesis, Stockholm University, Department of Linguistics, 2010.

Nuno Nobre. Anaphora resolution. Master's thesis, Instituto Superior Técnico/Universidade Técnica de Lisboa, 2011.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nate Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 492–501, Cambridge, MA, oct 2010. Association for Computational Linguistics.

Magnus Sahlgren. An Introduction to Random Indexing. *Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, TKE 2005, August 16*, 2005.

Sameer Singh, Amarnag Subramanya, Fernando Pereira, and Andrew McCallum. Large-scale cross-document coreference using distributed inference and hierarchical models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 793–803, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544, December 2001. ISSN 0891-2017.

Joel R. Tetreault. Analysis of syntax-based pronoun resolution methods. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 602–605. Association for Computational Linguistics, 1999.

A Targets and expansions

A note concerning the targets and expansions: When retrieval of targets has been performed, a regular expression without word boundaries has been used, thus retrieving genitive forms of different shapes. When retrieving expansions, a word boundary restricted the search in order to prevent overmatching.

A.1 Fredrik Reinfeldt

- Target strings

```
fredrik reinfeld
fredrik reinfelt
minister reinfeld
minister reinfelt
moderatledaren
partiledare reinfeld
partiledare reinfelt
statminister reinfeld
statminister reinfelt
statsminister reinfelt
```

- Expansion strings *lastname*

```
reinfeld
reinfeld's
reinfeld:s
reinfelds
reinfeldt
reinfeldt's
reinfeldt:s
reinfeldts
reinfelts
reinfelt's
reinfelt:s
reinfelt
```

- Expansion strings *firstname*

```
fredrik
fredrik's
```

```
fredrik:s
fredriks
```

- Expansion strings *epithet*

```
partiledare
statsminister
statsministern's
statsministern:s
statsministerns
```

- Expansion strings *pronoun*

```
han
hans
honom
```

A.2 Ron Paul

- Target strings

```
ron paul
ronald paul
ron ernest paul
ronald ernest paul
```

- Expansion strings *lastname*

```
paul
paul's
```

- Expansion strings *firstname*

```
ron
ron's
ronald
ronald's
```

- Expansion strings *epithet*

```
congressman
congressman's
republican
republican's
dr
dr\\.
doctor
doctor's
chairman
chairman's
```

- Expansion strings *pronoun*

```
he
his
him
```

A.3 Hugo Chávez

- Target strings

```
hugo chavez
hugo chaves
hugo chávez
hugo cháves
hugo rafael chavez
hugo rafael chaves
hugo rafael chávez
hugo rafael chávez
presidente chavez
presidente chaves
presidente chávez
presidente cháves
```

- Expansion strings *lastname*

```
chavez
chávez
chaves
cháves
```

- Expansion strings *firstname*

```
hugo
hugo rafael
```

- Expansion strings *epithet*

```
presidente
venezolano
comandante
coronel
lider
líder
jefe de estado
jéfe de estado
```

- Expansion strings *pronoun*

su
él
le