Coreference Resolution on Blogs and Commented News

Iris Hendrickx[1] and Veronique Hoste[1,2]

[1] LT3 - Language and Translation Technology Team,
University College Ghent,
Groot-Brittanniëlaan 45, Ghent,
Belgium
iris.hendrickx@hogent.be, veronique.hoste@hogent.be
[2] Department of Applied Mathematics and Computer Science,
Ghent University,
Krijgslaan 281(S9), Ghent,
Belgium
ve.hoste@ugent.be

Abstract. We focus on automatic coreference resolution for blogs and news articles with user comments as part of a project on opinion mining. We aim to study the effect of the genre shift from edited structured newspaper text to unedited, unstructured blog data. We compare our coreference resolution system on three data sets: newspaper articles, mixed newspaper articles and reader comments, and blog data. As can be expected the performance of the automatic coreference resolution system drops drastically when tested on unedited text. We describe the characteristics of the different data sets and we examine the typical errors made by the resolution system.

1 Introduction

One of the major challenges in an ever more globalizing world, in which the rise of the internet has led to a tremendous information and opinion overload, is the development of techniques which can assist humans in managing and exploiting this information wealth. Whereas, until recently, the international natural language processing research community mainly focused on the “factual” aspects of content analysis , we can observe a growing interest in the analysis of attitude and affect in textual sources. As messages (consumer reviews, blogs, e-mails, short messages, etc.) are becoming more prevalent on the Internet than edited (newswire) texts, it becomes crucial to develop robust technologies to extract not only the factual information, but also opinions, evaluations, beliefs and speculations from text.

In blogs, opinion sites, message boards, chats and forums, people can describe their personal experiences and opinions on about anything. People write about their personal life and express their opinions through writing blogs; they actively participate in discussions around the news by participating in forums or by posting comments on texts written by others. Newspapers have engaged in these trends: they no longer just publish their news articles online, but they offer their readers the opportunity to participate and publish their own comments and opinions about an article. News is also much more interactive as it is not published once a day as is the case with printed newspapers, but news stories are updated every time an event evolves. In case of major events, some newspapers even start live blogs offering people direct communication with the journalists present at the scene of the event.

As people are so productive in expressing their opinions on the web nowadays, their generated content is not only useful for anyone who has to make everyday decisions (which brand to choose, which movie to go to, which hotel to choose). Companies as well are anxious to understand how their services and products are perceived. Given the enormous amount of potentially interesting information, which is impossible to handle manually by media analysts, an automatic procedure is required which offers a digest of opinions on a certain product, service or company. This media reviewing procedure creates a variety of opportunities for individuals and organizations: to support companies in product and service benchmarking, to support market and competitor intelligence, in customer complaint management, in customer relation management, in advertising (associate advertisements with user-generated content), as decision support for political organizations, etc.

In order to support media analysts in their analysis of trends and opinions, automatic extraction tools are needed which are able to reliably detect the three basic components of an opinion [14]: (i) an *opinion holder*, viz. the person, institution, government, etc. that holds a specific opinion on a particular object, (ii) the *target*, i.e. a product, person, event, organization, topic, or even an opinion on which an opinion is expressed [20] and an *opinion* i.e. a view, attitude, or appraisal on an object from an opinion holder. The opinion classification could be tertiary (sentiment polarity classification) [28] or scalable (sentiment strength detection). Both the identification of the opinion holder and the target involve coreference resolution [21].

Coreferential resolution between the mentioned entities in the text and across different texts plays an important role in automatic opinion mining. We focus on automatic coreference resolution for blogs and news articles with user comments as part of a project on opinion mining for Dutch. We aim to study the effect of the genre shift from edited structured newspaper text to unedited, unstructured blog data. We compare our coreference system on three data sets: newspaper articles, mixed newspaper articles and reader comments, and blog data. Blogs can be seen an online diaries expressing the personal opinions of the blog author. They are often written in a style that resembles spoken language. Published news articles on the other hand are highly structured, factual and edited. On the point of referring expressions, blogs contain much more personal pronouns than newspaper text [16].

In the next Section we first describe related work on coreference resolution and opinion mining. Section 3 gives a detailed overview of the three data sets we use and describes the characteristics of the different text genres. In Section 4, we explain our experimental setup. Section 5 presents our results which are further discussed in this Section. Section 6 presents some concluding remarks.

2 Related work

Nicolov et al. [18] investigated the effect of coreference resolution for the task of product opinion mining in blog data. As text from a blog often contains topic drifts, they propose to use snippets of texts around a product name instead of full blog posts as a starting point for opinion extraction. In their study, they showed that information on coreference relations can improve their opinion mining system with approximately 10%.
The work of Stoyanov and Cardie [21] studies coreference resolution for opinion summarization. The authors focus on identifying opinion holders and resolving coreference relations between them. They work with partially annotated data in which only the opinion holder's coreferential information is annotated. They propose a new algorithm that can handle partially supervised clustering of this type of data. Choi et al. [4] and Bethard et al. [2] present closely related work, yet they aim at another type of relations. They study the recognition of entities and the relations between opinion holders and entities that themselves represent opinions or beliefs. According to Kobayashi et al. [13], opinion mining and anaphora resolution can be considered as a similar type of tasks: one can view linking an opinion to a source as linking an anaphor to an antecedent.

From a methodological point of view, coreference resolution on blog data could also benefit from prior work on coreference in dialogue. [23] describe a machine learning approach to the resolution of third person pronouns in spoken dialogue which uses a set of additional features which are specifically designed to handle spoken dialogue data (e.g. type of antecedent, verb's preference for arguments of a particular type). Their results show that these additional features are mainly beneficial for recall. [11] describe a rule-based system for handling anaphora in multi-person dialogues. The system integrates different constraints and heuristics, some of which are tailored to dialogues, but they do not evaluate the added value of these specific constraints and heuristics. [15] focus on coreference resolution in conversational documents (2007 ACE data) which incorporate speaker and turn information. They propose to use this metadata information to compute a group of binary features and show that this metadata information improves the ACE-value for broadcast conversation and telephone conversation documents. Given the (highly) unstructured nature of both dialogues and blogs, the insights from coreference

Table 1. Data statistics: number of tokens, sentences and average sentence length per data set.

Test set	# documents	# Tokens	# Sent.	Av. sent. length
Published news texts	25	111,117	576	19.3
News and comments	5	14,276	937	15.2
Blogs	15	5,689	289	19.7

resolution on dialogue data can be useful for coreference resolution on blogs. Our present study, however, is mainly focused on investigating the effect of genre shift; in the near future, we plan to investigate feature construction typically tailored to blog texts.

3 Data

In the present study, we aim to investigate the effect of the genre shift from edited structured newspaper text to unedited, unstructured blog data. In order to do so, we compared our coreference system on three data sets, namely newspaper articles, mixed newspaper articles and reader comments, and blog data:

As data set of *published news text* we used the KNACK 2002 data set which contains 267 Dutch news articles manually annotated with part-of-speech, named entities and coreferential information between noun phrases [9]. In the experiments presented here, we only use the manually annotated coreference links. For part-of-speech tags and named entities we use automatically predicted labels produced by automatic taggers as detailed in Section 4.

In WordNet 3.0 [7] a *blog* is defined as "a shared on-line journal where people can post diary entries about their personal experiences and hobbies; postings on a blog are usually in chronological order". A corpus of blogs has typical characteristics in terms of its content, structure and temporal aspects [16]. The author of a blog writes about his or her personal life often addressing many diverse topics and expresses individual comments, ideas and thoughts. The internal structure of a blog is a series of pieces of texts (posts). Timelines are an important feature of blogs as each post in the blog has a time stamp and the most recent posts are listed first. Blogs should not be seen as personal, isolated generated content, but rather as part of a network: blog posts contain links to other pages and many blogs offer readers the possibility to post reactions, making a blog interactive. As blogs are not edited they contain more spelling errors, ungrammatical sentences, and they deviate from newspaper text in terms of the use of capitalization, abbreviations and punctuation marks denoting emphasis or emoticons (like :D) or duration effects (e.g. ...).

Example 1 (Excerpt from news comments. Each comment has an author and time stamp.).

```
wtf is twitter
Drinkyoghurt | 31-03-09 | 00:35
---
Duh, its just texting  to a site so your friends can read
em there.  What'dya mean detour?
And sorry if I explain it wrong, that's cuz I don't give
a shit.
Ozdorp | 31-03-09 | 00:52
---
Those extra hours of training at school makes  teenagers
smarter apparently....
Paramada | 31-03-09 | 01:23
---
Twitter  doesnt stand a chance. They offer the same functio-
nality as  SMS (with respect to character limitation) plus
some functions that the rest of the internet (Google, Digg,
RSS) deals with in a much better way.  If you want to be
popular, do it with a suitable media method like Hyves.
(Not a fan either but I do have an account to get rid of
all that ridiculous 'JOIN HYVES' spam.)
Canterwood | 31-03-09 | 16:52
---
```

Our third source of data consists of *newspaper articles and reader comments* and is a mixture of text produced by professional writers and user-generated unedited text. The reader comments have the form of posts with a time stamp and are mostly displayed in chronological order. Both types of text address the same topic, but differ highly in style and are opposites in many perspectives such as formal versus informal, factual versus personal, edited versus unedited. Contrary to most blog posts which usually address all kinds of topics and thoughts, the reader comments of a news article have a focused topic. The posted reactions to news articles on news source websites have the same informal writing style and structural characterics as the blog data.

As an evaluation set, we collected 5 news articles with reader comments from an online newspaper and 15 blog posts. These were also manually annotated with coreferential information. The blog posts were collected from two blogs on Belgian cities and are written by multiple authors.The content of the blog posts varies from personal stories about a certain event to more informative blog posts describing upcoming events in the city.

We selected five news articles an accompanying comments. The selected news articles themselves are rather short, no longer than 20 sentences. The number of reader comments per article ranges from 88 to 123 different comments. In general these comments are short, the majority contains at most one or two sentences. The language use strongly resembles chat or spoken language. As an example of this type of data, we translated a excerpt of the comments on a Dutch news article stating that adolescents are not enthusiastic about Twitter shown in example 1. We consider each news article and the accompanying reader comments as one single document. This is a practical choice, many of the comments refer to the entities mentioned in the news article. However we do notice that our single document view is somewhat simplistic and not all characteristics of the data are well captured in our representation.

Table 1 gives an overview of the size of the different test sets. It mainly reveals that there are no differences in sentence length between the fairly structured blog data and the published news texts. The data set with the newspaper articles and reader comments, however, contains shorter sentences. Table 2 presents information about the type and quantity of anaphors in the different test sets. Our observations confirm the findings published in [16]; the blogs and commented news both contain relatively more pronouns. Here we focus on a quantitative overview of the number of pronouns which are not part of a coreference chain, presents a similar tendency: 61% of the pronouns in the data set containing the newspaper articles with reader comments does not refer to a preceding antecedent, whereas this percentage is much lower for the other two data sets (Published: 32.1% and Blogs: 38.7%).

Table 2. Proportion of pronominal, common noun and proper noun coreferential NPs. Number of pronouns which are not part of a coreference chain.

Test set	No coreference	Coreference			
	Pronouns	Pronouns	Proper N.	Common N.	All
Published News texts	178	282	426	492	1200
News and comments	610	390	200	537	996
Blogs	101	214	100	269	583

4 Experimental setup

The coreference resolution system takes a machine learning approach following the example of a.o. Soon et al. [19], Ng and Cardie [17] and is based on previous work of Hoste [10] for Dutch. Coreference resolution is seen as a classification task in which each pair of noun phrases in a text is classified as having a coreferential relation or not. For each pair of noun phrases, a feature vector is created denoting the characteristics of the pair of noun phrases and their relation.

To create the feature vectors, we first process the text. First, tokenisation is performed by a rule-based system using regular expressions. Part-of-speech tagging and text chunking is performed by the memory-based tagger MBT [5]. For the grammatical relation finding which determines which chunk has which grammatical relation to which verbal chunk (e.g. subject, object, etc.) a memory-based relation finder is used [24]. We also use a automatic Named Entity Recognition system, MBT trained on Dutch data set of the CoNNL 2002 shared task [25]. Besides these predicted labels (persons, organizations, locations, miscellaneous names), the system performs a look up names in gazetteer lists to supplement the automatic system, and to refine the predicted label *person* to *female* or *male*.
Several information sources contribute to a correct resolution of coreferential relations: morphological, lexical, syntactic, semantic and positional information and also world knowledge. In order to come to a correct resolution of coreferential relations, existing systems, e.g. [8, 3, 19, 22], use a combination of these information sources. For our coreference resolution system, we extract the following types of features: string overlap, distance between the noun phrases, overlap in grammatical role and named entity type, synonym/hypernym relation lookup in WordNet, morphological suffix information and local context of each of the noun phrases. For a more detailed description of the feature construction, we refer to [10].

We train different systems for different types of referring expressions. This allows us to optimize the system for each type of expression separately. Furthermore, splitting the treatment of the expressions can also help to focus on the errors separately for each referring expression made by the resolution system. We create three separate systems for pronouns, named entities and common nouns and optimize the machine learning classifier for each type separately. As machine learning algorithm we used memory-based learning as implemented in the software package Timbl [6]. We optimized the algorithmic parameters and feature weighting for each for each system with an heuristic search method that iteratively tries to find an optimal parameter setting for the data set at hand [26].

The experiments on the three different data sets are set up in the following way. We split the KNACK data set into a training set of 242 articles and a held out set of 25 articles for testing. The blog data set and news comments data set were only used for testing and not for training. We train our coreference resolution system on the KNACK training data and test it on each of the three different test sets. We measure the performance of our system using the MUC [27] and the B-Cubed [1] scoring software.

5 Results

We present the results of our coreference resolution on the three data sets in table 5. We computed precision, recall and F-score using the

MUC scoring and recall computed with the B-cubed method. As can be expected, the performance of the coreference resolution systems drops significantly for the blog and news with comments test sets. The results on the blog material are the lowest. The MUC scores and B-cubed scores show the same tendencies.

Table 3. Results of the coreference resolution system on the three different data sets: edited news paper text, blog data and news with reader comments. Scores computed with the MUC and B-cubed scoring methods.

	MUC scoring			B-cubed
Test set	recall	precision	F-score	recall
Published News texts	44.7	66.8	53.6	52.3
Blogs	18.9	40.0	25.7	43.5
News and comments	26.7	42.7	32.8	48.7

5.1 Error analysis

On the basis of a shallow manual error analysis on three texts of each corpus, we were able to detect typical errors that are made on the different data sets. The most problematic classes are the following:

- **Pronouns erroneously being classified as coreferential**: for the published news paper texts, we could observe a large number of pleonastic pronouns which were linked with a preceding noun phrase. The pleonastic pronoun was always the neutral 3 person singular pronoun “het”. The news and reader comments data set reveals the same tendency, but in this data set it is not restricted to the neutral 3 person singular pronoun. Also personal pronouns like “je” (you) or “zij” (they) are often used when referring to people in general and not to a specific entity mentioned in the text. e.g.

  (1) Du: Als **het** met dat coördinatiecentrum slecht afloopt (...)
  (En: If it doesn’t end well with that coordination centre...)

- **Incomplete detection of noun phrases**: all data sets share the problem of the incomplete detection of noun phrases which leads to partial detection of coreferential relations. e.g. in sentence 2 below, only part of the NP is recognized, viz. “Mevrouw”.

  (2) Du: **Mevrouw** Spiritus Dasesse zet heel geëmancipeerd haar meisjesnaam voorop (...)
  (En: Mrs Spiritus Dasesse puts her maiden name first)

- **Problems with the current feature vector**: for all data sets, the feature vector sometimes does not provide enough disambiguating information to distinguish between a positive and negative classification. e.g. in example 3, “elkaar” is erroneously linked to “de 190 miljoen euro”. The feature vector below illustrates the feature vector which was used as the basis for the positive classification.

(3) Du: Hij herhaalt dus alweer dat hij tegen half januari **de 190 miljoen euro** bij **elkaar** heeft (...)
(En: He repeats again that he'll have the 190 milioen euro by mid-January)
(7 519 1088) (elkaar) (7 518 1083) (de 190 miljoen euro) 0 1 miljoen euro bij TW(hoofd,prenom,stan) N(soort,ev,basis,zijd, stan) VZ(init) heeft om in WW(pv,tgw,met-t) VZ(init) VZ(init) dist_lt_two appo_no jpron_yes 0 0 0 0 num_na 0 0 0 0 0 0 0 0 I-OBJ I-OBJ I-OBJ person 0 0 0 0 0 zijdig 3p refl def_yes 0 0 0 0 NEG POS

- **Errors that need ’world knowledge’ or sophisticated information resources**: For some of the coreferential links a specialized resource such as an ontology or a database with gathered facts is needed to resolve the ambiguity. The abbreviation “MP” (minister president) in example 4 refers to earlier mentions in the text like “Balkenende” and “JPB”. To resolve these coreferential links one needs to know the name of the current minister president of the Netherlands. Our training material is not helpful because it is older than the comments news articles and blogs, so the names referring to ’minister president’ in the training material are different than the ones in this test material.

(4) Du: Het is in Nederland een grote rotzooi en **onze MP** maar praten over normen en waarden.
(En: The Netherlands is a big mess and our MP just talks about values.)

6 Conclusion

The work presented here can be seen as a first step towards a automatic coreference resolution system that will be integrated in a online automatic extraction tool for media analysis. Here we focused on examining the differences in language use between texts from (printed) news papers and mixed newspaper articles and reader comments and blog data. We studied the characteristics of the three different data sets in section 3. We experimented with an automatic coreference resolution system trained on edited news paper text and compared it’s performance on the three different text types. As expected, our results show that the performance of our automatic coreference resolution system drops significantly when confronted with unedited text. Next we examined in more detail the type of errors made by the system and the possible causes of these errors.

An obvious method to improve the coreference resolution system is to train not only on news paper articles, but also on a data set consisting of spoken language or annotated blogs and commented news data. However, we believe that adding training material will not be sufficient to resolve all problems. In an adapted version of our coreference system we also plan to add additional features. We would like to add factual information gathered from the web or from available corpora. Finding facts is a method that is regularly applied in question-answering systems e.g. [12]. This type of information can be seen as a resource of 'world knowledge' and help to resolve ambiguities like the one illustrated in example 4.
The discussion on related work on dialogues already suggested that information on turn taking can be valuable. We expect this to be true for blogs and reader comments as well. Especially for pronouns in the commented news data set, explicit information about turn taking can help our system to resolve pronouns that refer to the author or to authors of previous comments. Because our system already has a separate trained module for pronominal anaphors, it will be relatively easy to adjust the system on this point.

Acknowledgements

The work presented here was conducted within the framework of the DuOMAn (Dutch Online Media Analysis) project which is funded by the Dutch-Flemish STEVIN research program. We would like to thank the anonymous reviewers for their helpful comments.

References

1. Amit Bagga and Breck Baldwin. Algorithms for scoring coreference chains. In *Proceedings of the First International Conference on Language Resources and Evaluation Workshop on Linguistic Coreference*, pages 563–566, 1998.
2. Steven Bethard, Hong Yu, Ashley Thornton, Vasileios Hatzivassiloglou, and Dan Jurafsky. Automatic extraction of opinion propositions and their holders. In *In 2004 AAAI Spring Symposium on Exploring Attitude and Affect in Text*, pages 22–24, 2004.
3. C. Cardie and K. Wagstaff. Noun phrase coreference as clustering. In *Proceedings of the 1999 joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 82–89, 1999.
4. Yejin Choi, Eric Breck, and Claire Cardie. Joint extraction of entities and relations for opinion recognition. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2006.
5. W. Daelemans, J. Zavrel, A. van den Bosch, and K. van der Sloot. Memory based tagger, version 2.0, reference guide. Technical Report ILK Technical Report - ILK 03-13, Tilburg University, 2003.

6. W. Daelemans, J. Zavrel, K. Van der Sloot, and A. Van den Bosch. TiMBL: Tilburg Memory Based Learner, version 6.1, reference manual. Technical Report 07-07, ILK, Tilburg University, 2007.
7. C. Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.
8. F. Fisher, S. Soderland, J. Mccarthy, F. Feng, and W. Lehnert. Description of the umass system as used for muc-6. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pages 127–140, 1995.
9. V. Hoste and G de Pauw. Knack-2002: a richly annotated corpus of dutch written text. In *The fifth international conference on Language Resources and Evaluation (LREC)*, 2006.
10. Véronique Hoste. *Optimization Issues in Machine Learning of Coreference Resolution*. PhD thesis, Antwerp University, 2005.
11. Prateek Jain, Manav Ratan Mital, Sumit Kumar, Amitabha Mukerjee, and Achla M. Raina. Anaphora resolution in multi-person dialogues. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, 2004.
12. Valentin Jijkoun, Maarten De Rijke, and Jori Mur. Information extraction for question answering: Improving recall through syntactic patterns. In *In Coling 2004*, pages 1284–1290, 2004.
13. Nozomi Kobayashi, Ryu Iida, Kentaro Inui, and Yuji Matsumoto. Opinion extraction using a learning-based anaphora resolution technique. In *Second International Joint Conference on Natural Language Processing: Companion Volume including Posters/Demos and tutorial abstracts*, pages 175–180, 2005.
14. B. Liu. *Web Data Mining. Exploring Hyperlinks, Contents and Usage Data*. Springer, 2006.
15. Xiaoqiang Luo, Radu Florian, and Todd Ward. Improving coreference resolution by using conversational metadata. In *Proceedings of NAACL HLT 2009*, pages 201–204, 2009.
16. Gilad Misnhe. *Applied Text Analytics for Blogs*. PhD thesis, University of Amsterdam, Amsterdam, The Netherlands, 2007.
17. V. Ng and C. Cardie. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002)*, pages 104–111, 2002.
18. N. Nicolov, F. Salvetti, and S. Ivanova. Sentiment analysis: Does coreference matter? In *Proceedings of the Symposium on Affective Language in Human and Machine*, Aberdeen, UK, 2008.
19. W. M. Soon, H. T. Ng, and D. C. Y. Lim. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544, 2001.
20. V. Stoyanov and C. Cardie. Topic identification for fine-grained opinion analysis. In *Proceedings of the Conference on Computational Linguistics (COLING 2008)*, 2008.
21. Veselin Stoyanov and Claire Cardie. Partially supervised coreference resolution for opinion summarization through structured rule learning. In *Proceedings of the Conference on Empirical Methods on Natural Language Processing (EMNLP)*, pages 336–344 336–344. Association for Computational Linguistics, 2006.

22. M. Strube, S. Rapp, and C. Müller. The influence of minimum edit distance on reference resolution. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP-2002)*, pages 312–319, 2002.
23. Michael Strube and Christoph Müller. A machine learning approach to pronoun resolution in spoken dialogue. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 168–175, 2003.
24. E.F. Tjong Kim Sang, W. Daelemans, and A. Höthker. Reduction of dutch sentences for automatic subtitling. In *Computational Linguistics in the Netherlands 2003. Selected Papers from the Fourteenth CLIN Meeting*, pages 109–123, 2004.
25. Erik F. Tjong Kim Sang. Introduction to the conll-2002 shared task: Language-independent named entity recognition. In Dan Roth and Antal van den Bosch, editors, *Proceedings of CoNLL-2002*, pages 155–158, 2002.
26. A. van den Bosch. Wrapped progressive sampling search for optimizing learning algorithm parameters. In *Proceedings of the 16th Belgian-Dutch Conference on Artificial Intelligence*, pages 219–226, 2004.
27. M. Vilain, J. Burger, J. Aberdeen, D. Connolly, and L. Hirschman. A model-theoretic coreference scoring scheme. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pages 45–52, 1995.
28. T. Wilson, J. Wiebe, and P Hoffman. Recognizing contextual polarity: an exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, 2008.