

**INTERNATIONAL WORKSHOP**  
**EVENTS IN EMERGING TEXT TYPES (eETTs)**

*held in conjunction with the International Conference*  
*RANLP - 2009, 14-16 September 2009, Borovets, Bulgaria*

**PROCEEDINGS**

Edited by  
Constantin Orăsan, Laura Hasler and Corina Forăscu

Borovets, Bulgaria

17 September 2009

**International Workshop**

**EVENTS IN EMERGING TEXT TYPES (eETTs)**

**PROCEEDINGS**

Borovets, Bulgaria  
17 September 2009

ISBN 978-954-452-011-3

Designed and Printed by INCOMA Ltd.  
Shoumen, Bulgaria

## Foreword

The proliferation of the Internet has revolutionised the way information is disseminated and presented. Blogs no longer just relay and comment on news stories but also influence what is talked about in the news. Such changes have not gone unnoticed by the computational linguistics research community, which is increasingly processing or exploiting blogs in an attempt to keep track of what is going on and mine information. This workshop focuses on how events can be identified and how information related to event processing (e.g. NP coreference, temporal processing) can be extracted from blogs and other online sources. Emphasis is on how existing methods for event processing need to be adapted in order to process this medium, and on linguistic differences in the reporting of events in blogs and more traditional news texts.

Event detection and processing is not a new topic in computational linguistics, but until now it has focused mainly on processing of newswire. The TimeBank corpus (Pustejovsky et. al. 2003), the AQUAINT TimeML corpus, and the NP4E corpus (Hasler, Orasan and Naumann 2006) exclusively contain newswire, which may make them inappropriate for the development of methods which need to process other text types. Moreover, the informal style and structure of most blog entries makes event detection in these documents a difficult task. This workshop gives researchers the opportunity to present efforts to develop resources related to event identification and processing using blog entries, including annotation guidelines and linguistic analyses of such resources.

We would like to thank the organisers of the 7th International Conference on Recent Advances in Natural Language Processing, RANLP 2009, for hosting this workshop as one of their satellite events. We would also like to thank the researchers who submitted papers and the members of Programme Committee for their help in the reviewing process.

We hope you enjoy the workshop,

The organising committee



## **Organisers and Sponsors**

**Events in Emerging Text Types (eETTs) is organised by:**

**Constantin Orăsan**, University of Wolverhampton, UK

**Laura Hasler**, University of Wolverhampton, UK

**Corina Forăscu**, “A.I. Cuza” University of Iași, Romania



## PROGRAMME COMMITTEE

**Alexandra Balahur**, Alicante University, Spain  
**Carmen Banea**, University of North Texas, USA  
**Sabin-Corneliu Buraga**, “A.I. Cuza” University of Iași, Romania  
**Di Cai**, University of Wolverhampton, UK  
**Andras Csomai**, Google  
**Iustin Dornescu**, University of Wolverhampton, UK  
**Erhard Hinrichs**, Tübingen University, Germany  
**Véronique Hoste**, University College Gent, Belgium  
**Radu Ion**, Research Institute for AI, Romanian Academy  
**Rafael Muñoz**, University of Alicante, Spain  
**Vivi Năstase**, EML Research, Germany  
**Nicolas Nicolov**, Umbria Inc, USA  
**Georgios Paltoglou**, University of Wolverhampton, UK  
**Livia Polanyi**, Powerset/Microsoft, USA  
**Irina Prodanof**, ILC-CNR, Pisa, Italy  
**Mark Rogers**, Market Sentinel, UK  
**Doina Tătar**, “Babes-Bolyai” University, Romania





## Table of Contents

<i>Summarizing Blog Entries versus News Texts</i> Shamima Mithun and Leila Kosseim .....	1
<i>Event Detection in Blogs using Temporal Random Indexing</i> David Jurgens and Keith Stevens .....	9
<i>A Pairwise Event Coreference Model, Feature Impact and Evaluation for Event Coreference Resolution</i> Zheng Chen, Heng Ji and Robert Haralick .....	17
<i>Summarizing Threads in Blogs Using Opinion Polarity</i> Alexandra Balahur, Elena Lloret, Ester Boldrini, Andrés Montoyo, Manuel Palomar and Patricio Martínez-Barco .....	23
<i>Catching the news: two key cases from today</i> Ruslana Margova and Irina Temnikova .....	32
<i>Detecting Opinion Sentences Specific to Product Features in Customer Reviews using Typed Dependency Relations</i> Ashequl Qadir .....	38



# Workshop Program

**Friday, September 17, 2009**

## **Session 1**

- 9:20–9:30 Welcome and opening remarks
- 9:30–10:30 Invited Talk by Mark Rogers: *The value of language processing to commercial clients*
- 10:30–11:00 *A Pairwise Event Coreference Model, Feature Impact and Evaluation for Event Coreference Resolution*  
Zheng Chen, Heng Ji and Robert Haralick

## **Session 2**

- 11:30–12:00 *Catching the news: two key cases from today*  
Ruslana Margova and Irina Temnikova
- 12:00–12:30 *Summarizing Blog Entries versus News Texts*  
Shamima Mithun and Leila Kosseim

## **Session 3**

- 14:00–15:00 Invited Talk by Mike Thelwall: *Detecting public news interests from blogs and MySpace*
- 15:00–15:30 *Summarizing Threads in Blogs Using Opinion Polarity*  
Alexandra Balahur, Elena Lloret, Ester Boldrini, Andrés Montoyo, Manuel Palomar and Patricio Martínez-Barco

## **Session 4**

- 16:00–16:30 *Detecting Opinion Sentences Specific to Product Features in Customer Reviews using Typed Dependency Relations*  
Ashequl Qadir
- 16:30–17:00 *Event Detection in Blogs using Temporal Random Indexing*  
David Jurgens and Keith Stevens
- 17:00–17:30 Round table/Closing session



# Summarizing Blog Entries versus News Texts

Shamima Mithun and Leila Kosseim  
Concordia University

Department of Computer Science and Software Engineering  
Montreal, Quebec, Canada  
{*s\_mithun, kosseim*}@encs.concordia.ca

## Abstract

As more and more people are expressing their opinions on the web in the form of weblogs (or blogs), research on the blogosphere is gaining popularity. As the outcome of this research, different natural language tools such as query-based opinion summarizers have been developed to mine and organize opinions on a particular event or entity in blog entries. However, the variety of blog posts and the informal style and structure of blog entries pose many difficulties for these natural language tools. In this paper, we identify and categorize errors which typically occur in opinion summarization from blog entries and compare blog entry summaries with traditional news text summaries based on these error types to quantify the differences between these two genres of texts for the purpose of summarization. For evaluation, we used summaries from participating systems of the TAC 2008 opinion summarization track and updated summarization track. Our results show that some errors are much more frequent to blog entries (e.g. topic irrelevant information) compared to news texts; while other error types, such as content overlap, seem to be comparable. These findings can be used to prioritize these error types and give clear indications as to where we should put effort to improve blog summarization.

## Keywords

Opinion summarization, blog summarization, news text summarization.

## 1 Introduction

Everyday, people express their opinions on a variety of topics ranging from politics, movies, music to newly launched products on the web in weblogs (or blogs), wikis, online-forums, review sites, and social networking web sites. As more and more people are expressing their opinions on the web, the Internet is becoming a popular and dynamic source of opinions. Natural language tools for automatically mining and organizing these opinions on various events will be very useful for individuals, organizations, and governments.

Various natural language tools to process and utilize event-related information from texts have

already been developed. Event-based question answering systems [21] and event-based summarization systems [12] are only a few examples. However, most of the event-based systems have been developed to process events from traditional news texts. Blog entries are different in style and structure compared to news texts. As a result, successful natural language approaches that deal with news texts might not be as successful for processing blog entries; thus adaptation of existing successful NLP approaches for news texts to process blog entries is an interesting and challenging task. The first step towards this adaptation is to identify the differences between these two textual genres in order to develop approaches to handle this new genre of texts (blogs) with greater accuracy. In this study, we compare automatically generated summaries of blog entries with summaries of news texts with the goal of improving opinion summarization from blog entries. In particular, we compared summaries for these two genres of texts on the basis of various errors which typically occur in summarization.

In this paper, we first investigate what kind of errors typically occur in query-based opinionated summary for blog entries. The errors that we have identified are categorized and then used to compare blog summaries with news texts summaries. For evaluation, we used summaries from participating systems at the TAC 2008 [1] opinion summarization track and updated summarization track. Summaries of the TAC 2008 opinion summarization track and updated summarization track were generated from blogs entries and traditional news texts, respectively. The systems participating in the TAC opinion summarization track and in the updated summarization track are quite different in several aspects, as they are targeted to resolve two different tasks. The systems participating in the updated summarization track were mainly required to find the answers to given queries and detect redundant information while the systems participating in the opinion summarization track were required to perform opinion mining and polarity classification in addition. Moreover, the systems participating in the opinion summarization track were provided optional snippets (described in section 3.1) and were restricted with a maximum summary length which were much higher compared to the updated summarization track. Despite these differences, these two datasets were used in our work because they are the most comparable datasets for our task.

## 2 Characteristics of Blogs

Blogs (or weblogs) are online diaries that appear in chronological order. Blogs reflect personal thinking and feelings on all kinds of topics including day to day activities of bloggers; hence an essential feature of blogs is their subjectivity. Some blogs focus on a specific topic while others cover several topics; some describe personal daily lives of bloggers while others describe common artifacts or news. Many different sub-genres of blogs exist. The two most common are personal journals and notebooks [5]. Personal journals discuss internal experiences and personal lives of bloggers and tend to be short [5]. They are usually informal in nature and written in casual and informal language. They may contain much and sometimes only unrelated information such as ads, photos, and other non-textual elements. They also contain spelling and grammatical errors, and punctuation and capitalization are often missing. On the other hand, notebooks contain comments on internal and external events. Similarly to newspaper articles, they are usually long and written in a more formal style [5]. Most NLP work on blogs has tended to study personal journals as opposed to notebooks. For example, the Blog-06 corpus [15], used at TREC and at TAC, contains mostly personal journals.

## 3 Blog Summarization

Opinion summarization, and in particular blog summarization, is a fairly recent field. Some systems (e.g. [9, 10]) have been developed for opinion summarization to generate a summary from a document. In 2008, the Text Analysis Conference (TAC) introduced a query-based opinion summarization track. They provided questions, a blog corpus and optional snippets which are found by QA systems. These query-based summarization systems are designed to retrieve specific answers on an event or entity instead of an overview of the whole document.

Opinion summarization uses opinionated documents such as blogs, reviews, newspaper editorials or letters to the editor to answer opinionated questions. On the other hand, summarization of traditional news texts uses fact-based information such as formal and non-opinionated texts. As we are interested in opinion summarization from blog entries, we will use the two terms *opinion summarization* and *blog summarization* interchangeably.

### 3.1 Current Approaches

A query-based opinion summarizer recapitulates what people think or feel on a particular topic (or an event or entity) by answering a specific query. For example, one such opinionated query could be *What has been Russia's reaction to U.S. bombing of Kosovo?*. A query-based opinion summarizer can answer opinion questions posed in natural language; thus it helps users to get specific answers to questions they are interested in, instead of retrieving an entire document.

At the TAC 2008 opinion summarization track, a set of target topics on various events or entities were

given on which participating systems were evaluated. For each topic, a set of questions and a set of relevant blog entries (mostly personal journals) were provided. For example, for the topic “*UN Commission on Human Rights*”, two questions were asked:

1. “*What reasons are given as examples of their ineffectiveness?*”
2. “*What steps are being suggested to correct this problem?*”

and a set of IDs of related blog entries were provided. Systems needed to extract answers to questions from these specified sets of blog entries. Additionally, some sample answer snippets were provided for every topic that summarization systems may use. These snippets were extracted by the participating QA systems at the TAC 2008 QA track. Here are two sample snippets for the topic *UN Commission on Human Rights*:

1. “*Issues regular resolutions condemning Israel while overlooking real offenders.*”
2. “*To ensure this new body would be no facsimile of its predecessor, the legislation prohibits membership to countries that violate human rights or are subject to specific human rights resolutions.*”

Two types of summarization approaches were used by TAC participants, namely: snippet-driven approaches and snippet-free approaches. Snippet-driven approaches use snippet information to extract sentences which contain these snippets from the input blog entries. They then generate a summary by incorporating these sentences. Snippet-free approaches do not use snippets. They mainly utilize query information and sentiment degree for sentence scoring. Participating systems first filter blog entries to identify the relevant content and remove irrelevant information such as ads, photos, music, videos, and other non-textual elements. The focus and polarity of the question are identified; then sentences are ranked according to their relevance with the query. The polarity of the sentences is also calculated and matched with the polarity of the query. To find the relevance with the query, overlap with the query terms is calculated using different techniques such as the cosine similarity, language models etc. Opinion dictionaries and different machine learning techniques are used to identify the polarity of the question and sentences. Finally, the summaries are generated using the ranked sentences.

### 3.2 Evaluation

Evaluation of blog summaries use the same criteria as for traditional news text summarization. The quality of a summary is assessed mostly on its content and linguistic quality [14]. Content evaluation of a query-based summary is performed based on the relevance assessment (with the topic and query) and inclusion of important contents from the input documents.

Currently, the automatic evaluation tool ROUGE [11] is the most popular evaluation approach for content evaluation. ROUGE automatically compares system generated summaries with a set of

model summaries (human generated) by computing n-gram word overlaps between them. Conferences and workshops such as TAC and DUC (Document Understanding Conference) [2] use ROUGE. The pyramid method [18] is also used for content evaluation. In the pyramid method, multiple human generated summaries are analyzed manually to generate a gold standard. In this process, summary analysis is done semantically such that information with the same meaning (expressed using different wording) is marked as summary content unit (SCU). A weight is assigned for each SCU based on the number of human summarizers that express it in their summaries. In this method, the pyramid score for a system generated summary is calculated as follows [17]:

$$\text{score} = (\text{the sum of weights of SCUs expressed in a generated summary}) / (\text{the sum of weights of an ideally informative summary with the same number of SCUs})$$

The linguistic quality of a summary is evaluated manually based on how it structures and presents the contents. Grammaticality, non-redundancy, referential clarity, focus, structure and coherence are the commonly used factors considered to evaluate the linguistic quality. Mainly, subjective evaluation is done to assess the linguistic quality of an automatically generated summary. In this process, human assessors directly assign scores on a scale based on agreement or disagreement with predefined set of questions such as “Are they ungrammatical?”, “Do they contain redundant information?”. The assessments are done without reference to any model summaries.

### 3.3 News Text Summarization versus Blog Summarization

As most work has been done on news text summarization, it is not surprising that the performance of such systems are generally higher than blog summarizers. For example, as shown in Table 1, at the TAC-2008 conference, the average scores for news text summaries (updated summarization track) are higher than for blog summaries (opinion summarization track) using all 3 evaluation criteria.

**Table 1:** Average TAC-2008 Summarization Results - Blogs vs. News Texts

Genre	Pyramid Score	Linguistic Score	Resp. Score
Blogs	0.21	2.13	1.61
News	0.27	2.33	2.32

Table 1 shows summary evaluation using the pyramid score, linguistic quality and responsiveness (Resp.). The last two criteria were evaluated by human assessors on a scale of 1 to 5 (1, being the worst). In this evaluation, the responsiveness of a summary was judged to measure the overall quality or usefulness of the summary, considering both the

information content and readability.

This difference in performance between blogs and news texts can be attributed to the differences in the two textual genres. Indeed, one of the essential characteristics of blogs as opposed to news texts, is their subjectivity (or opinion). Unlike traditional news text summarization, sentiment (subjectivity) plays a key role for blog summarization. For blog summarization, sentiment degree is often used to rank sentences. In the case of query-based blog summarization, the sentiment polarity of the question needs to be matched with that of summary sentences.

In addition, as opposed to traditional news texts, blogs are usually written in casual language. For blogs, it is usually very difficult to identify which portions of blog entries are relevant to the topic. News texts are more uniform in style and structure. Blogs may contain many unrelated information such as ads, photos, music, videos. For blogs, it is often difficult to find sentence boundaries. In most cases punctuation and capitalization are unreliable. As a result, for blog summarization, systems need to put additional efforts to pre-process the text compared to news text summarization. Furthermore, because blogs do not exhibit a stereotypical structure, some features such as position of sentence, or similarity with the first sentence, which are shown to be useful for traditional news text summarization are not as useful for blog summarization [6].

## 4 Error Analysis

To identify the different challenges posed by blog summarization as opposed to traditional news texts summarization, we have studied 50 summaries from participating systems at the TAC 2008 opinion summarization track and compared these to 50 summaries from the TAC 2008 updated summarization tracks. The average summary length of the opinion summarization track was 1224 words, while that of the updated summarization track was 179 words. The average input documents length of the opinion summarization track was 1888 words, while that of the updated summarization track was 505 words. Summaries were randomly selected for the evaluation. However, we ensured that we selected summaries from all participating systems on all topics. The task of the updated summarization track was chosen for comparison because it is similar in nature to the blog summarization task in the sense that its goal is also to generate query focused (but non-opinionated) summaries (using news articles). Even though there are several differences between the summarization approaches in TAC opinion summarization track and updated summarization track, these two datasets are the most comparable datasets for our task.

In this study, we have analyzed the most common types of errors in our 100-summary corpus and have categorized them in 3 main categories:

1. *Summary-Level Error (SuLE)*
2. *Sentence-Level Error (SeLE)*



### 3. Intra-Sentence-Level Error (ISLE)

These are shown in Figure 1 and discussed in the following sub-sections.

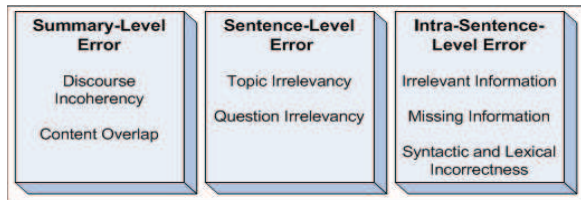


Fig. 1: Types of Errors in Blog vs. News Summaries

#### 4.1 Summary-Level Errors

We define a Summary-Level Error (SuLE) as the textual contents which reduce the understandability and readability of the overall summary. There are two types of SuLE:

1. Discourse Incoherency (DI)
2. Content Overlap (CO)

##### Discourse Incoherency (DI)

A summary will exhibit a Discourse Incoherency (DI) if the reader cannot identify the communicative intentions of the writer from the propositions or if the propositions do not seem to be interrelated [8]. In the sample summaries that we have studied, Discourse Incoherency occurred both at the sentence level and at the proposition level.

Consider the following summary (ID:T1004.20 <sup>1</sup>) where a DI occurs at the sentence level:

**Topic:** Starbucks coffee shops

**Question:** Why do people like Starbucks better than Dunkin Donuts?

**Summary:** I am firmly in the Dunkin' Donuts camp. It's a smooth, soothing cuppa, with no disastrous gastric side effects, very comforting indeed. I have a special relationship with the lovely people who work in the Dunkin' Donuts in the Harvard Square T Station in Cambridge. I was away yesterday and did not know. [...]

In this summary, the underlined sentence is not coherent with the rest of the text because it is not interrelated with the rest of the text.

Consider the following summary (ID:T1001.3) where a DI occurs at the proposition level:

**Topic:** Carmax

**Question:** What motivated positive opinions of Carmax from car buyers?

**Summary:** At Carmax, the price is the price and when you want a car you go get one. Tyler Sink Says: December 7th, 2005 at 9:22 am,

<sup>1</sup> All summaries numbered ID:Txxxx.xx are taken from the TAC 2008 opinion summarization track.

with Carmax you will generally always pay more than from going to a good used car dealer. [...]

In the underlined sentence, the textual fragments *Tyler Sink ..* and *with Carmax ..* do not seem to be related to one another.

##### Content Overlap (CO)

If semantically similar information exists in different units of a text, then we define it as Content Overlap (CO). Content overlap can range from a simple duplication of text fragments to a more complex textual entailment problem. For example, consider the summary below (ID:T1019.35):

**Topic:** China one-child per family law

**Question:** What complaints are made about China's one-child per family law?

**Summary:** [...] If you have money to pay the fines, you can have 2 or 4 children. [...] \$6400 - a typical fine for having more than one child- in China is about 2-3 years salary. [...] Imagine losing your job, being fined 2-3 years salary for having a second child. [...]

In this summary, the underlined sentences carry similar contents. So it may seem redundant to include all these sentences in the final summary.

Table 2: Summary-Level Errors - Blogs vs. News Texts

Error Type		Blogs	News	$\Delta$
DI	Discourse Incoherency	30.44%	10.66%	19.78%
CO	Content Overlap	19.14%	14.66%	4.48%

Table 2 compares Summary-Level errors in our 50 blog summaries corpus and our 50 news texts summaries corpus. Table 2 shows that opinionated blog summarization and non-opinionated news texts summarization both exhibit an important number of *Discourse Incoherency* and *Content Overlap* errors. However, blog summarization have around 20% more *Discourse Incoherency* and about 4.5% more *Content Overlap* errors, than those of news article summarization. We suspect that the reason behind this is that blogs are generally informal in nature. As a result, in blogs, propositions are often incoherent and contain redundant information. On the other hand, the formal nature of news articles reduces these errors for news texts summarization.

#### 4.2 Sentence-Level Errors

If a summary sentence is irrelevant to the central topic of the input documents or to user query, then the summary contains a Sentence-Level Error (SeLE). Two types of SeLE were identified:

1. Topic Irrelevancy (TI)



## 2. Question Irrelevancy (QI).

### Topic Irrelevancy (TI)

As mentioned in Sections 3.1 and 4, in both the TAC 2008 opinion summarization track (blogs) and the updated summarization track (news texts), participating systems needed to generate a summary answering a set of questions on a specific target (topic). However, in both tasks, many systems generated a summary containing sentences that are not related to the specified topic. Here is an example of a TI (ID:T1004.33):

**Topic:** *Starbucks coffee shops*

**Question:** *Why do people like Starbucks better than Dunkin' Donuts?*

**Summary:** *Well ... I really only have two. [...] I didn't get a chance to go ice-skating at Frog Pond like I wanted but I did get a chance to go to the IMAX theatre again where I saw a movie about the Tour de France it wasn't that good. [...]*

### Question Irrelevancy (QI)

Many of the system generated summary sentences are not relevant to the question even though they are related to the topic. An example of a QI is shown below (ID:T1004.3):

**Topic:** *Starbucks coffee shops*

**Question:** *Why do people like Starbucks better than Dunkin' Donuts?*

**Summary:** *Posted by: Ian Palmer — November 22, 2005 at 05:44 PM Strangely enough, I read a few months back of a coffee taste test where Dunkin' Donuts coffee tested better than Starbucks. [...] Not having a Dunkin' Donuts in Sinless City I am obviously missing out... but Starbucks are doing a Christmas Open House today where you can turn up for a free coffee. [...]*

The underlined sentence is relevant to the topic but not to the question.

**Table 3:** *Sentence-Level Errors - Blogs vs. News Texts*

Error Type		Blog	News	$\Delta$
TI	Topic Irrelevancy	41.67%	5.86%	35.81%
QI	Question Irrelevancy	47.87%	16.67%	31.20%

Table 3 compares Sentence-Level errors for blog summaries and news text summaries. Note that in the table, *Topic Irrelevancy* is calculated based on the entire corpus. However, *Question Irrelevancy* is calculated based only on the sentences which are related to the topic. Table 3 shows that a large number of sentences from blog summaries have *Topic Irrelevancy* and *Question Irrelevancy* errors. In contrast, in news text summarization, *Topic Irrelevancy* error

occurs only occasionally and *Question Irrelevancy* error is also not very frequent. Blogs summarization has around 30% more of these two errors than that of news text summarization. We suspect that the main reason behind such a difference is brought about by the summary evaluation scheme. Indeed, many systems use the optimal summary length (7000 characters per question) allowed in TAC which results in many out of context sentences to be used as filler. As a result, the average summary length of the opinion summarization track is much longer than that of the updated summarization track (1224 words versus 179 words). Another important reason for these errors is the informal style and structure of blog entries. Indeed, sentences in blog entries do not have a predictable rhetorical structure (e.g. in formal writing, the first and the last sentences of a paragraph usually contain important information) which can be used to rank sentence during summarization. As a result, it is much more difficult to rank blog sentences compared to news text sentences. Opinion (sentiment) information is typically used to rank blog sentences for summarization. We also believe that because opinion identification can be quite imprecise, it can possibly add more noise to the blog sentence ranking process. Moreover, unlike pre-focused news articles, blogs are quite unfocused. In blogs, bloggers express various opinions about the topic which are not relevant to the question. Together all these issues may lead to a high number of topic and question unrelated sentences in blog summarization.

## 4.3 Intra-Sentence-Level Errors

Intra-Sentence-Level (ISLE) errors occur within a sentence and involve irrelevant or missing information, grammatical errors, or lexical errors (e.g. spelling errors). Intra-Sentence-Level Errors include:

1. *Irrelevant Information (II)*
2. *Missing Information (MI)*
3. *Syntactic and Lexical Incorrectness (SLI)*

Each of these categories are described below with examples.

### Irrelevant Information (II)

Under Irrelevant Information (II) errors, a significant portion of a sentence is irrelevant to the summary topic or question. For example, consider the summary below (ID:T1003.9):

**Topic:** *Jiffy Lube*

**Question:** *What reasons are given for liking the services provided by Jiffy Lube?*

**Summary:** *They know it's fine cause Jiffy Lube sent them a little card in the mail and they have about a month before they need an oil change. [...] Well, they suppose it is a little bit of a PITA to figure out what to do with the spent oil, but after some digging, they found out that every Jiffy Lube will take used oil for free! [...]*

The underlined snippet above is irrelevant to the question even though it holds a coherent discourse relation with the last proposition.

### Missing Information (MI)

If a sentence does not contain all the necessary information to make it comprehensible for the reader and the required information to understand the sentence is also not available in the context then this error is defined as a Missing Information (MI) error.

Here is an example of MI. In the following summary (ID:T1021.17):

**Topic:** *Sheep and Wool Festival*

**Question:** *Why do people like to go to Sheep and Wool festivals?*

**Summary:** *[...] i hope to go again this year and possibly meet some other knit bloggers this time around since i missed tons of people last year. I love going because of the tons of wonderful people, yarn, Sheep, rabbits, alpacas, llamas, cheese, sheepdogs, fun stuff to buy, etc. , etc. [...]*

The underlined sentence contains incomplete information, which cannot be resolved from the context either making it incomprehensible.

### Syntactic and Lexical Incorrectness (SLI)

Syntactical level errors such as grammatical incorrectness and incompleteness of a sentence or lexical level errors such as spelling errors, short forms, stylistic twists of informal writing ... in a sentence is defined as Syntactic and Lexical Incorrectness (SLI) error.

For example, consider the following summary (ID:T1009.32):

**Topic:** *Architecture of Frank Gehry*

**Question:** *What compliments are made concerning his structures?*

**Summary:** *Central to Millennium Park in Chicago is the Frank Gehry-designed Jay Pritzker Pavilion, described as the most sophisticated outdoor concert venue of its kind in the United States. [...] Designing a right-angles-be-damned concert hall for Springfield, hometown of Bart et al.. [...]*

In this summary, the underlined sentence is an example of a SLI.

**Table 4:** *Intra-Sentence-Level Errors - Blogs vs. News Texts*

Error Type		Blog	News	$\Delta$
II	Irrelevant Information	30.91%	15.66%	15.25%
MI	Missing Information	9.33%	2.33%	7.00%
SLI	Syntactic and Lexical Incorrectness	18.79%	4.00%	14.79%

Table 4 compares Intra-Sentence-Level errors for blog summaries and news text summaries. From Table 4, we can see that *Irrelevant Information*, *Missing Information*, and *Syntactic and Lexical Incorrectness* errors appear about 15%, 7%, and 15% more respectively in blog summarization. Here again, we believe that the informal nature of blogs explains these difference.

## 5 Discussion

Compared to a manual linguistic evaluation of a summary, our work tries to identify and quantify the differences in error types between two textual genres: blogs and news.

Our error types incorporate both what the automatic and manual summary evaluation try to measure. Indeed, Sentence-Level Errors (Topic Irrelevancy and Question Irrelevancy) evaluate the content and relevance of the summaries similarly to what ROUGE tries to evaluate; whereas the remaining errors (Summary-Level Errors and Intra-Sentence Errors) evaluate more the linguistic quality of a summary.

It is not surprising to see that Topic Irrelevancy, Question Irrelevancy, Discourse Incoherency, Irrelevant Information and Syntactic and Lexical Incorrectness are much more frequent in blogs than in news texts (from 36% to 19% more frequent). Content Overlap and Missing Information, on the other hand, seem to be only slightly more frequent (5% and 7%) in blogs summaries than in news texts summaries. These results give a clear idea of how difficult it is to process blog entries for summarization compared to news texts and where efforts should be made to improve such summaries.

## 6 Related Work

### 6.1 NLP on blogs

Recently, the availability of opinions on current events on weblogs opened up new directions in natural language research. Even though natural language processing on blogs is a fairly new trend, its popularity is growing rapidly. Many conferences and workshops (e.g. [1, 3, 4, 15]) are taking place to address different aspects of the analysis of blog entries. Current NLP work on blog entries include: subjectivity and sentiment analysis; question answering; and opinion summarization.

Subjectivity and sentiment analysis include classifying sentiments of reviews [19] and analyzing blogger mood and sentiment on various events [16]. Sentiment classification of reviews on different events is often done on movie or product reviews. Rating indicators of reviews are used to identify the polarity of the blogs namely positive, negative or neutral. To analyze blogger mood and sentiment, systems make use of information regarding bloggers' mood varying over time. To record bloggers' varying mood, the polarity information of the blog post is often used. Some works (e.g. [16]) are done to measure how bloggers' varying mood affects different events. In addition, the TREC

blog track [15] provides an opportunity to build new techniques of sentiment tagging on blog posts. The task is to identify and rank blog posts on a given topic from a corpus of blog entries.

Question answering (QA) on blog entries is a relatively new field. Most notable QA work on blog entries was conducted at TREC 2007 [15] and TAC 2008 [1]. To answer queries on an event or entity, TREC provided a blog corpus in addition to the AQUAINT newspaper corpus [15].

## 6.2 Analysis of blogs versus news

To the best of our knowledge there have been only a few work carried out to compare the difference between blog entries and news texts; however, none seems to have analyzed it at the linguistic level for a specific NLP application.

Ku et al. [10] developed a language independent opinion summarization approach. For summarization, they retrieved all sentences which are relevant to the main topic of the document set and determined the opinion polarity and degree of these relevant sentences. They also found that the identification of correlated events on a time interval is also important for opinion summarization. They tested their approach for blog entries and news texts for English and Chinese languages. From their evaluation, they found that blog entries contain more topic irrelevant information compared to news texts. Their results confirm our own results. Ku et al. also found that news texts use a larger vocabulary compared to blog entries which makes the filtering of non-relevant sentences task harder for news texts. On the other hand, this larger vocabulary helps to decide sentiment polarities. Due to the limited vocabulary the judgment of sentiment polarity of blog entries was difficult.

Somasundaran et al. [20] developed an opinion question answering approach for blogs and news texts. They exploited attitude information namely sentiment and argument types to answer opinion questions. They received comparable result with both text types.

Lloyd et al. [13] developed the Lydia system to analyze blog entries. They analyzed temporal relationship between blogs and news texts. In particular, they analyzed how often bloggers report a story before newspapers and how often bloggers react to news that has already been reported. To study this leads/lag relationship, they analyzed frequency time series of 197 most popular entries in news texts and blog corpora over six week period. Lydia first recognized name entities to extract information from both corpora. Then the system resolved noun phrase coreference because a single entity is often mentioned using multiple variations on their name. Then it performed a temporal analysis to identify which entities are referred more frequently over a certain period of time. In their analysis, they found that 30 entities exhibited no lead/lag relationship, 73 had news leading the blogs, and 94 had blogs leading the news.

Godbole et al. [7] developed a large-scale sentiment analysis system on top of the Lydia text analysis system [13] for news texts and blog entities. They determined the public sentiment on various entities

and identified how this sentiment varies with time. They found that the same entities (person) except certain controversial political figures received comparable opinions (favorable or adverse) in blogs and news texts. Controversial political figures received different opinions in blogs compared to news texts because of the political biases among bloggers, and perhaps the mainstream press.

Though both the work Lloyd et al. and Godbole et al. handle news text and blog entries, their application domains (temporal relationship and sentiment analysis) are different from ours. Somasundaran et al. tested their question answering approach for news texts and blogs. They compared their approach for both genres of text mainly on the basis of subjectivity information. On the other hand, we compared summaries of both text types on the basis of errors which mainly occurred from informal style and structure of blog entries. Our work is most similar to Ku et al.'s work. However, we identified a larger number of errors of summarization and compared blog summaries with traditional news texts summaries on the basis of these errors. As a result, our work will better enable us to pinpoint the difference between these two genres of texts for summarization task.

## 7 Conclusion

As the performance of blog summarization is generally much lower than for news text summarization, we set out to compare automatically generated summaries for blogs entries with news texts based on the most common errors which occurred in summarization. The goal of our comparison was to assess whether these summary related errors affect traditional news texts based non-opinionated summaries differently than opinionated blog summaries.

We first analyzed and categorized errors that occur in opinion summarization on blogs using the summaries from participating systems at the TAC 2008 opinion summarization track. Then we compared these results with those of the TAC 2008 updated summarization track. Our results show that all types of summary related errors occur more often in blog summarization than news texts summarization. However, topic and question irrelevancy pose a much greater problem for blog summarization than for traditional news texts; while content overlap and missing information seem to be only slightly more frequent in blog than traditional news texts. These findings can be used to prioritize these error types and give clear indications as to where we should put effort to improve blog summarization.

## 8 Acknowledgements

The authors would like to thank the anonymous referees for their valuable comments on an earlier version of the paper.

This work was financially supported by NSERC.

## References

- [1] Text Analysis Conference (TAC): <http://www.nist.gov/tac>. (Last accessed 2009-05-20).
- [2] Document Understanding Conferences (DUC): <http://duc.nist.gov>. (Last accessed 2009-05-20).
- [3] Third International AAAI Conference on Weblogs and Social Media, San Jose, California, May 2009.
- [4] Third Annual Workshop on the Blogging Ecosystem: Aggregation, Analysis, and Dynamics. In *Workshop Of WWW-2006*, Edinburgh, May 2006.
- [5] A. Andreevskaia, S. Bergler, and M. Urseau. All Blogs are Not Made Equal: Exploring Genre Differences in Sentiment Tagging of Blogs. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM-2007)*, Boulder, Colorado, March 2007.
- [6] A. Bossard and M. Genereux. Description of the LIPN Systems at TAC 2008: Summarizing Information and Opinions. In *Notebook Papers and Results, Text Analysis Conference (TAC-2008)*, Gaithersburg, Maryland, USA, November 2008.
- [7] N. Godbole, M. Srinivasaiah, and S. Skiena. Large-Scale Sentiment Analysis for News and Blogs. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM'2007)*, pages 219–222, Boulder, Colorado, USA, March 2007.
- [8] E. H. Hovy. Automated Discourse Generation using Discourse Structure Relations. *Artificial Intelligence*, 63(1-2):341–385, 1993.
- [9] M. Hu and B. Liu. Mining and Summarizing Customer Reviews. In *SIGKDD 2004*, pages 168–177, 2004.
- [10] L. W. Ku, Y. T. Liang, and H. H. Chen. Opinion Extraction, Summarization and Tracking in News and Blog Corpora. In *Proceedings of the AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs*, California, USA, March 2006.
- [11] C. Y. Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain, July 2004.
- [12] M. Liu, W. Li, M. Wu, and H. Hu. Event-Based Extractive Summarization using Event Semantic Relevance from External Linguistic Resource. In *Proceedings of the Sixth International Conference on Advanced Language Processing and Web Information Technology, ALPIT 2007*, pages 117–122, Henan, China, 2007.
- [13] L. Lloyd, P. Kaulgud, and S. Skiena. Newspapers vs. Blogs: Who Gets the Scoop? In *Proceedings of the AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs, 2006*, California, USA, March 2006.
- [14] A. Louis and A. Nenkova. Automatic Summary Evaluation without Human Models. In *Notebook Papers and Results, Text Analysis Conference (TAC-2008)*, Gaithersburg, Maryland (USA), November 2008.
- [15] C. Macdonald, I. Ounis, and I. Soboroff. Overview of the TREC 2007 Blog Track. In *Proceedings of the Sixteenth Text REtrieval Conference (TREC 2007)*, Gaithersburg, Maryland, USA, November 2007.
- [16] G. Mishne and N. Glance. Predicting Movie Sales from Blogger Sentiment. In *Proceedings of the AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW 2006)*, 2006.
- [17] A. Nenkova. Summarization Evaluation for Text and Speech: Issues and Approaches. In *Proceedings of Interspeech 2006*, Pittsburg, USA, 2006.
- [18] A. Nenkova and R. Passonneau. Evaluating Content Selection in Summarization: The Pyramid Method. In *Proceedings of the HLT/NAACL, 2004*.
- [19] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment Classification using Machine Learning Techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2002.
- [20] S. Somasundaran, T. Wilson, J. Wiebe, and V. Stoyanov. QA with Attitude: Exploiting Opinion Type Analysis for Improving Question Answering in On-line Discussions and the News. In *Proceedings of the International Conference on Weblogs and Social Media*, Boulder, Colorado, USA, March 2007.
- [21] H. Yang, T. S. Chua, S. Wang, and C. K. Koh. Structured use of External Knowledge for Event-based Open Domain Question Answering. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 33–40, Toronto, Canada, 2003.



# Event Detection in Blogs using Temporal Random Indexing

**David Jurgens and Keith Stevens**  
University of California, Los Angeles  
4732 Boelter Hall  
Los Angeles, CA 90095  
{jurgens, kstevens}@cs.ucla.edu

## Abstract

Automatic event detection aims to identify novel, interesting topics as they are published online. While existing algorithms for event detection have focused on newswire releases, we examine how event detection can work on less structured corpora of blogs. The proliferation of blogs and other forms of self-published media have given rise to an ever-growing corpus of news, commentary and opinion texts. Blogs offer a major advantage for event detection as their content may be rapidly updated. However, blogs texts also pose a significant challenge in that the described events may be less easy to detect given the variety of topics, writing styles and possible author biases. We propose a new way of detecting events in this media by looking for changes in word semantics. We first outline a new algorithm that makes use of a temporally-annotated semantic space for tracking how words change semantics. Then we demonstrate how identified changes could be used to detect new events and their associated blog entries.

## 1 Introduction

Automated event detection is a form of information retrieval where given a time-ordered set of documents, an algorithm must select those which represent recent news or changes to existing information. An automated approach to event detection has many practical applications; given the large amount of text being written daily, readers want to be informed of which developments and topics are the most recent and important without having to manually sift through all the documents written on the topic. In addition, a robust system should be able to detect multiple kinds of events, such as international conflicts, product releases or sports results. The main challenge in automating this task is detecting what makes a new document sufficiently novel to be described as a new event.

Current event detection approaches have focused on identifying concrete events that occur within newswire text[11]. However, in recent years, blogs have become an important source of both news and commentary. Unlike news reports, blog content expresses a wide range of topics, opinions, vocabulary and writing styles; the change in editorial requirements allows blog authors to comment freely on local, national and international issues, while still expressing their personal sentiment. Accordingly, blogs offer a rich opportunity for detecting events that may not be covered in traditional newswire text. These forms of self-published media might also allow event detection systems to identify developing events before official news reports can be written.

Several forms of event detection have focused on analyzing named entities, such as “Bill Clinton” or “Iraq,” and the contexts or documents in which they appear, e.g. [10, 11, 5]. We propose a more general approach that looks at all words and their contexts, rather than a predetermined set of words. Specifically, we argue that event detection can be done by measuring the semantic change in a word or phrase. To track changes in the semantics, we use a semantic space model of meaning, which is an automated method of building distributed representations of word meaning.

Semantic space models of meaning offer three notable advantages for event detection. First, the models are capable of automatically determining the semantics of a word by examining the contexts in which the word appears. Such automated understanding of semantics is required for analyzing these new sources of data due to the much wider vocabulary used by authors. Second, the models offer a well defined method for comparing the semantics between words. These semantic comparisons have been shown to be similar to human judgments[13]. We argue that reporting words which have a notable changes in semantics should correlate well with a reader’s expectations of interesting developments. Third, the models are well-established at detecting association such as synonymy among words, which can allow models to detect events that are referred to by multiple names. Given these advantages, we introduce a new semantic space algorithm for assessing how the meaning of a word changes through time for the purpose of event detection.

We illustrate our approach to topic detection with a hypothetical example of the product release of a toy named “blick.” At the start of the toy’s popularity, the word “blick” has not occurred before and therefore its semantics would be undefined. As “blick” appears in more blogs, the word acquires consistent semantics, and the algorithm can report a new event for “blick.” Our approach differs from simple occurrence monitoring in that we require the word to have a consistent meaning; unless the algorithm is capable of determining what concepts the word refers to, knowing that the word relates to an event is impossible.

However, consider detecting a second event for “blick” soon after its release in which the toy is discovered to have toxic properties. Since the toy’s name was already present in the blogs, the novelty of the name is not enough to detect the point at which the toxic chemical was revealed. However, our approach, which looks at the semantic shift of words over time, would detect a shift based on the new kinds of words that would be likely to co-occur with the toy’s name, e.g. toxicity, a toy recall, or lawsuit. Intuitively speaking, this approach associates news events with noticeable changes in both *what* authors talk about and *how* they

talk about those subject.

In this paper we present a new algorithm, Temporal Random Indexing, that effectively captures the semantic changes for words and phrases over time. We first briefly review the semantic space model that underlies this approach and then present the algorithm. Following, we demonstrate several examples of semantic change extracted from a large blog corpus and illustrate one method for reporting the events.

## 2 Semantic Space Models

Semantic space models of meaning are born from the distributed hypothesis: For two words, their similarity in meaning is predicted by the similarity of their distributions of co-occurring words[6], or as Firth puts it, “you shall know a word by the company it keeps,”[4]. Creating semantics from co-occurring words forms the basis for how our algorithm represents changes in semantics.

### 2.1 Semantics as Co-occurrence

In a semantic space, a word’s semantics are mapped to high dimensional vectors in a geometric space. The dimensions of the space represent distinctions between the meanings of words; accordingly, words with similar semantics have similar vector representations. Semantic space representations have proven effective at a variety of information retrieval tasks such as identifying synonymous queries[21] and multi-language retrieval[14, 23]. For a recent survey of applications of semantic spaces to information retrieval see Cohen and Widdows[3]. To illustrate the basics of co-occurrence based semantic space models, we can further explore the example of “blick”, the new yet toxic toy.

Consider the documents describing “blick” when it is first introduced during a holiday season. A potential line from several blogs might read “A perfect gift this holiday season is blick, one of the newest toys available!” Using a simple co-occurrence semantic space, the semantics of “blick” would be a count of how frequently it co-occurs with key words such as: gift, holiday, perfect and toys. Examining later blog posts written when this same toy is discovered to have toxic elements, several posts might now have the line: “the toxic elements in blick make the toy dangerous.” The semantics of the toy should now focus primarily on the co-occurrence of words such as toxic and dangerous, and should no longer be associated with positive words such as holiday and perfect. Figure 1 illustrates a simplified two-dimensional semantic space and the changes to semantics that would occur as “blick” begins to co-occur with toxic-related words. A standard semantic space model would define the semantics of the new toy as a combination of all co-occurrences, in this case the positive new semantics and the negative semantics of toxicity.

### 2.2 Random Indexing

Using simple co-occurrence is rarely done in practice for large corpora. In such models, each unique word would be assigned its own dimension (corresponding to co-occurrence with that word), which results in vectors with hundreds of thousands to millions of dimensions. Basing the number of dimensions on the number of unique words

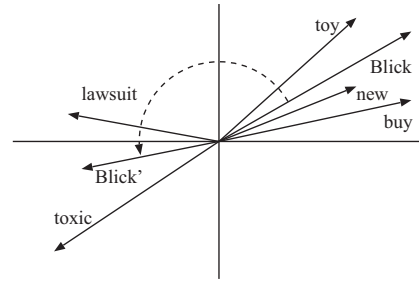


Fig. 1: Word semantics projected into two dimensions, illustrating the hypothetical change in meaning for “blick” based on its nearest neighbors

is particularly problematic for blog corpora, as writers frequently introduce misspellings, slang, or topic-specific jargon. Accordingly, many approaches have focused on reducing the dimensionality of the semantic space. Dimensionality reduction often has the additional benefits such as making the resulting vector more general, or reducing computation time.

Early successful approaches such as Latent Semantic Analysis[13] use the Singular Value Decomposition (SVD) to reduce the number of dimensions. While the SVD results in significant improvements in information retrieval, the fastest algorithms for the SVD are  $O(mn^2)$  [7], which make them impractical for large corpora. Moreover, the SVD and other forms of principle component analysis must have the entire corpus present at once, which makes it difficult to update the space as new words and contexts are added. This is particularly problematic for event detection, as the corpus is expected to continuously grow as new events occur.

Random Indexing[9, 18] offers an alternative method for reducing the dimensionality of the semantic space by using a random projection of the full co-occurrence matrix onto a lower dimensional space. Random Indexing operates as follows. Each unique word is assigned an *index vector*, which is a random, sparse vector in a high dimensional space, often 2000-10000 dimensions. The size of the index vectors sets the number of dimensions used in the resulting semantic space. Index vectors are created such that any two arbitrary index vectors have a high probability of being orthogonal. This property is necessary to accurately approximate the original word co-occurrence matrix in a lower dimension. The semantics of each word are calculated by summing the index vectors of all co-occurring words within a small window of text. Random Indexing works well in practice as the dimensionality reduction occurs as the corpus is being processed, rather than requiring an explicit step after all the corpus has been seen.

More formally, let  $w$  be a focus word,  $w_i$  be a co-occurring word with a word distance of  $i$  and  $index(w_i)$  be the co-occurring word’s index vector. For the current word, we define a window of size  $n$  words before and after, which are counted as co-occurring. The semantics of  $w$  are then defined as:

$$semantics(w) = \sum_{c \in D} \sum_{-n \leq i \leq n} index(w_i) \quad (1)$$

where  $c$  is each occurrence of  $w$  in the corpus  $D$ .

### 2.3 Adding Time to Semantic Space Models

Augmenting a semantic space with time has been recognized as an effective method for tracking changes in semantics[20]. Two methods have been used to add temporal semantics. The first approach builds a separate semantic space for each specific time range. Semantics are then compared across spaces by defining some common context which occurs in both spaces. The second approach builds a single semantic space but provides the ability to segment it based on time. The key difference between these approaches lies in the meaning of each semantic dimension; when multiple spaces are used, there is no guarantee that the specific semantic meaning associated with some dimension  $i$  will be the same for dimension  $i$  in another space.

Kontostathis et al.[10] and Fortuna et al.[5] have independently proposed two successful semantic space algorithms that use the first approach of processing several distinct corpora. Both approaches collect several corpora which span unique time ranges, and construct a semantic space for each corpus using LSA. Using LSA is a notable challenge as the space defined by LSA is based on the SVD of a word  $\times$  document matrix; with documents being unique to each time-span’s corpus, direct comparison of vectors between spaces is not feasible.

Kontostathis et al.[10] use data mining to overcome the change in dimension-meaning by first clustering the semantics from each year. With this clustering, key attributes are extracted from several time ranges, and significant differences are used to infer an event or trend. In essence, vector comparisons between the semantic spaces are bypassed by using cross-space meta-statistics for each word generated from each space. This approach is limited to being an offline approach due to the costly machine learning techniques, and is further limited by key sets of attributes.

Another approach for comparing semantics from semantic spaces has been introduced by Fortuna et al.[5]. Their approach focused on finding key words that existed in multiple spaces, and defining a concrete set of semantics for these landmark words. As semantics from distinct spaces are created, they can be evaluated according to their relation to these landmark terms, and at any point in time, the words most closely associated to the landmark provide terms describing events related to the landmarks.

Sagi et al. propose an alternate approach of uses a single corpus and includes temporal semantics after generating an initial set of semantics[17]. This generates semantic vectors for a corpus spanning many time ranges of interest and reducing dimensionality via SVD. Then, to develop temporal semantics for a term, documents from a specific time range are used to generate temporal vectors through a process very similar to Random Indexing; in this process the first set of semantic vectors generated are used in place of index vectors when using equation (1).

While these approaches allow for accurate representations of semantic shifts, they face significant challenges when scaling to a large streaming set of documents, due to a reliance on the SVD for dimensionality reduction. Additionally, none of the algorithms are able to change the time-spans used for analysis without reprocessing some portion of a corpus. Given these limitations, a computationally efficient modification to how a semantic space is produced is necessary to permit more detailed analysis of changing semantics.

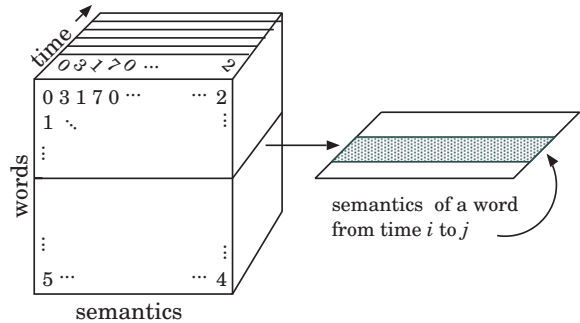


Fig. 2: The tensor representation of the semantic space

### 3 Temporal Random Indexing

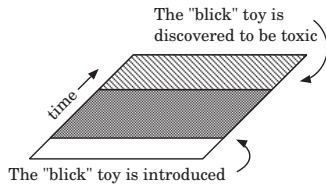
Temporal Random Indexing (TRI) incorporates time by building a single semantic space. However, instead of using a word  $\times$  semantics matrix, TRI uses a word  $\times$  semantics  $\times$  time tensor. Figure 2 illustrates this change. The time dimension records the semantic vectors of a word for each time unit (e.g. a week or month).

TRI offers three major advantages over existing models. First, the semantic space is built incrementally so new documents may be added at any time. Second, the tensor representation allows for arbitrary time-range comparisons. Third, we use a dimensionality reduction similar to that of Random Indexing, which operates as the documents are processed, rather than all at once. This results in greatly reduced time and memory requirements compared to those methods that rely on the SVD for dimensionality reduction.

To associate specific time values with semantics, TRI does not immediately perform the summation as defined in equation (1). Instead TRI only accumulates the semantics for contexts which occur in the same time period. These time period semantics are then stored in chronologically ascending order to produce a semantic slice for a word. The plane in figure 2 represents the semantic slice of a single word, which covers all the time periods in which that word has been observed.

Summing a semantic slice along the time dimension produces a vector equivalent to the results of Random Indexing, which would simply sum all the values, ignoring time, to create a single vector for the word. TRI, on the other hand, allows for more precise summations to be computed, such as a summation over the entire known time range, a single point in time, or several separate time ranges. Considering the example of a new toy that is first introduced and later found to be toxic, figure 3 shows how one could produce two semantic vectors for the toy “blick“ using a semantic slice. The first semantic vector is a summation of temporal semantic vectors that describe the introduction of the new toy, and the second semantic vector is a summation of temporal semantic vectors that describe the toxic nature of the toy. Using this technique, TRI can produce two distinct semantic representations of the same word, based on a simple partition of the temporal dimension. More over, these two vectors can be still be directly compared because they are built from the same index vectors.

Temporal Random Indexing can be formally described as a modification of equation (1), with three additional equations. As input, it takes an annotated collection of



**Fig. 3:** The semantic slice of “blick” as the meaning changes due to shifts in word co-occurrence patterns

documents  $D = ((t_0, d_0), (t_1, d_1), (t_2, d_2), \dots, (t_k, d_k))$ , where  $d_i$  is the set of documents occurring at time  $t_i$ . Let  $W_D$  be the set of all unique words in the collection. Just as in Random Indexing, for each word  $w \in W_D$ , we assign a unique index  $index(w)$ .

Equation(1) can then be extended to be:

$$semantics(w, t) = \sum_{c_t \in d_t} \sum_{-n \leq i \leq n} index(w_i) \quad (2)$$

where  $t$  is a unique timestamp, and  $c_t$  is the context for an occurrence of  $w$  at time  $t$ . Using this new definition of semantics, a word slice can be defined as:

$$slice(w) = \{(t_i, semantics(w, t_i)) | w \in d_i, i = 1, k\} \quad (3)$$

The semantics of a word for some range of time can then easily be computed with:

$$snapshot(w, t_i, t_j) = \sum_{(t_m, s_m) \in slice(w), t_i \leq t_m \leq t_j} s_m \quad (4)$$

## 4 Experiments

We applied TRI to the task of detecting events for a manually selected set of 199 words from a variety of topics. We selected words based on how frequently it was used in a corpus and knowledge that it would be likely to be discussed in blogs. However, limiting the word selection was done for efficiency, and this approach could be applied to tracking events for a larger set of words. Due to limited space, we illustrate the performance using a set of six word of divergent topics that includes both abstract and specific concepts: college, Lebanon, nuclear, Wii, PS3, and XP.

### 4.1 The Corpus

Our approach can be applied to any corpus that has a known date of authorship of each article, at the granularity desired for analysis of semantic shifts. For the purpose of detecting changes in public opinion over the course of recent events, and the detection of previously unknown, but still interesting, events, we have utilized a portion of an already existing corpus[22].

The corpus comes from a collection of blog postings from 2004 on. These blog postings come from around the world, and in a variety of languages. We view this as an excellent example of an unstructured corpus for event detection since it is composed of blog articles harvested by BlogLines<sup>1</sup>. The documents come from some standard

<sup>1</sup> <http://www.bloglines.com>

news sources, but also from any blogging service which provides rss feeds, such as livejournal, local newspapers, wordpress, and many more.

For this experiment, we collected only English articles from the blog corpus, but the algorithm could be used in practice with any language. The date of authorship for each document in this collection is estimated to be the most recent date the document has been updated.

Overall we expect this corpus to be well fitted to the challenge of detecting events while handling multiple view points beyond editorial control. Table 1 provides three sample blog posts which exemplify the issue. Each of the posts were written near the release of the Wii game console, each with a significantly different usage of words, and sentiment. There is a clear range of styles, from the mechanical description of the device, to opinions on the company releasing the system, and finally to adoration of the system. Beyond this sample set of posts, the corpus meets our expectations in other ways. First, the lack of editorial oversight in the documents leads to grammatical and spelling errors, and frequently to the introduction of new terms or phrases unique to the author along with other issues<sup>2</sup>. Second, the corpus has a large number of discussed topics, ranging from international events, to product releases, and to personal musings.

Before the corpus is used for performing event detection, the corpus is preprocessed to render it more uniform. Similar to other semantic space approaches that used web-gathered data[16], this pre-processing allows the model to gracefully handle several irregularities in writing style, such as inconsistent use of punctuation and capitalization. Additionally, this process removes many tokens such as html mark-up, which have little or no semantic content in themselves<sup>3</sup>. The corpus is processed as follows:

1. Replace all numbers with <num>
2. Remove all html mark-up and email addresses
3. Remove unusual punctuation, and separate all other punctuation from words
4. Remove words of 20 characters in length
5. Converting all words to lower case
6. Replacing \$5 to <num> dollars
7. Discard articles with fewer than some threshold percentage of correctly spelled English words
8. Associate each entry with a numeric timestamp

When computing the semantics, we also impose two filters on corpus during processing: any word in a list of frequent closed-class words and those words not in the most frequent 250,000 words in the blog corpus were removed. This step is both practical and empirically motivated.

Removing closed-class is a common practice in semantic spaces models[16, 19], due to the low semantic value; words such as “the” or “of” so frequently appear that they do not serve to distinguish the meaning of any co-occurring word. Similarly, infrequent words can safely be removed for initial uses due to the small effect they would have on other semantic vectors.

For both stop words and infrequent words, their original position is preserved after removal. This ensures that the window for counting co-occurrence takes into account

<sup>2</sup> This may lead to an increase in polysemy and synonymy amongst words, potentially impacting our approach, but exploration of this topic is left for future work

<sup>3</sup> We note that the HTML might be interpreted to yield more information however, TRI is agnostic to its input, and so no special HTML processing is done



For those of you who went, I hope you guys had just as much fun as I did. One of the best parts was actually being able to play the Wii. When I picked up that controller, I was sold instantly. The other awesome part was being able to demo for Enchanted Arms in the Ubisoft area. I have a bunch of pictures up on my Flickr Account if people are interested.	With motion-sensing controls and three-dimensional movements on screen, the upcoming Nintendo Wii game platform is changing the way video game developers think about games.	While I agree that expanding video games beyond its core audience is certainly an intriguing idea (if not necessary), it doesn't exactly thrill me as a member of said audience. Nintendo has already done a fine job of turning us all into their little marketing minions with the DS. None of this will change with the Wii. I call it exploitation of the weak spot in our hearts for the big N.
---	--	--

**Table 1:** *Contrasting blog entries about the Wii gaming console prior to its release in November 2006*

the words originally within the window distance. All remaining words and tokens are assigned an index vector for computing the semantics.

We limited the analysis to the 2006 postings in the corpus; this constituted 15,725,511 blog entries and a total of 2.62 billion tokens (both words and punctuation) after the normalization process.

## 4.2 Detecting Events using TRI

Events are extracted using a three step process. First, TRI is used to convert the corpus into semantic slices. A month-long time span was selected after an empirical analysis of the particular corpus showed that the reduced frequency of words in smaller time spans led to semantics that performed less well. TRI was configured using 10,000 dimensional vectors with a  $\pm 3$  word window. Index vectors had the values of 4 dimensions randomly assigned to +1 or -1, and the rest to be 0. Processing the entire corpus using TRI took approximately 100 minutes on a 2.4GHz Intel Core 2 processor with 8 gigabytes of RAM.

In the second step, the semantic shift is calculated for each word. To detect the shift, a word's semantic vectors for slices at time  $t_i$  and  $t_{i+1}$  are compared using the cosine similarity, which measures the similarity in angle between vectors. The cosine similarity ranges between 1 and -1, indicating identical and opposing angles, respectively. The semantic shift is defined as the  $1 - \text{cosine similarity}$ . Changes in angle reflect a change in a word's meaning, which in this system can signify the presence of an event. Changes in magnitude were also tracked but an analysis showed they were not correlated with events. Table 2 shows semantic shifts for several test words.

The third step selects those topic words that undergo a significant semantic shift and associate the topic words with documents. We define the significance for a shift in terms of its deviation from the mean semantic shift using a simple time series analysis. Specifically, we calculate the mean and standard deviation for the semantic shift of all words in the two slices. If a word's shift is greater than one standard deviation away from the mean, then it the word is marked as undergoing a significant shift. The bold values in table 2 note these shifts for five example words.

To form the association between documents and topic words, each word that undergoes a significant shift has its nearest neighbors calculated. These neighbors are often words associated with the topic word, but are not necessarily synonyms. We posit that the neighbors provide context about the nature of an event by virtue of reflecting the frequent co-occurrences in the documents. To retrieve the event-related documents, the topic word and its neighbors are used as query terms to search the corpus during the month that the event occurred. Documents are retrieved using a simple technique that returns the posts containing the

event term and the highest frequency of the related terms.

Accurately evaluating event detection requires a set of events that are known a priori to be in the corpus. For large corpora with millions of documents, such as the Bloglines corpus used here, it is infeasible to determine the complete set of events that are present. Furthermore, determining what kinds of events may be present can prove problematic, as blogs frequently discuss many topics outside the range of normal news events. To create a baseline for evaluation, We constructed a limited set of significant news events that were likely to be in the corpus and then manually verified their presence. Descriptive keywords for each event were then to evaluate TRI. Ultimately, the evaluation is an analysis of not only TRI, but also the corpus itself, as some terms, or events, may not be present at all within the corpus. We plan to use this initial methodology to identify a better means of analyzing massive corpora and the diverse set of events contained therein.

## 4.3 Results

Several semantic shifts correlated well with known events of 2006. We discuss the results by analyzing the events detected for the words in table 2. Table 3 lists some of the highest rated blogs associated with specific events our technique detected.

Both the "Wii" and the "PS3" are gaming consoles released in North America in November 2006. However, only the Wii experienced a significant semantic shift. The stabilization of the semantics correlates with the products demonstration at the Electronics Entertainment Expo, a major gaming event. The corpus contained many examples of attendees describing their experiences with both consoles at the convention. Notably the PS3 underwent only a slight shift, indicating a fairly stable meaning. Further analysis showed that the change in "Wii" was due to the console being renamed from "Revolution" to "Wii" in late April.

The 2006 Lebanon War took place in July 2006, which was detected by a significant shift in meaning and is further supported by a change in the nearest neighbors. In July, the nearest neighbors of "Lebanon" were terms associated with war, such as "Hezbollah", "soldiers", and "rockets". However, before, and after the war, the ten closest neighbors to "Lebanon" in 2006 were names of countries, revealing that during the course of the war, the semantics of "Lebanon" shifted dramatically to a different class of words, and then returned to its original class once the war concluded.

The changes for "nuclear" correspond directly to claims that North Korea conducted nuclear tests in October 2006. Until October, the related terms of "nuclear" are focused on Iran, and nuclear power; during October, the neighbors shift towards terms such as "Korea", "atomic", "sanctions", and "bomb."

	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
college	0.00	0.00	0.00	0.00	0.05	0.05	0.00	0.00	0.00	0.00	0.00
Lebanon	0.04	0.05	0.05	0.06	<b>0.16</b>	<b>0.25</b>	0.01	0.01	0.02	0.03	0.01
nuclear	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.01	<b>0.11</b>	0.06	0.02
PS3	0.06	0.04	0.06	0.05	0.03	0.04	0.03	0.02	0.02	0.03	0.01
Wii	<b>0.12</b>	<b>0.14</b>	<b>0.15</b>	0.06	0.03	0.04	0.04	0.02	0.01	0.01	0.01
XP	0.03	0.04	0.03	0.03	<b>0.19</b>	<b>0.20</b>	0.02	0.02	0.02	0.02	0.01

**Table 2:** Semantic shift values for six example words where bold indicates a significant change

Throughout the year, “college” experienced no noticeable semantic shifts, despite the annual events of beginning and graduating college. We view this as example of a consistent word which acts as a reference point to other words.

An analysis for “XP” showed a semantic shift caused by an unlikely change in corpus content; during the month of June, spammers added such a high number of advertisements for Windows XP copies, shifting the semantics to unimportant terms. An examination of the nearest neighbors to “XP” showed a dramatic change from related operating system terms such as “Windows,” “Linux” and “Vista” to numbers and currency abbreviations.

Overall, using the cosine similarity metric, and then further examining the sets of nearest neighbors proved to be an effective method of catching semantic shifts. Furthermore combining the nearest neighbors with a simple document retrieval algorithm generated relevant documents corresponding to the events which caused the shift.

## 5 Related Work

Among the many systems that perform event detection, several use a related technique that maps documents, rather than words, into a vector space. Documents are represented by a vector of their term frequencies, with most approaches using some form of weighting the vectors, such as term frequency-inverse document frequency (TF-IDF) weighting. Additionally, many event detection systems restrict themselves to processing newswire text, instead of blogs.

Most document based event detection algorithms extend a core usage of TF-IDF weighting. In the core event detection algorithm, each document is analyzed to produce TF-IDF values for each word occurring in the document. This set of values can then be compared against TF-IDF values of other documents using the same similarity measures used in semantic space models, with cosine similarity being one of the most common. In general, an event is detected if the current document is significantly different from all other processed documents, based on a threshold of similarity values between documents[1, 11].

Brants et al. introduced some significant improvements to the standard document based model[1]. The first improvement was to compute the TF-IDF values on a document by document basis, allowing the system to continuously process new documents. The second improvement was computing a set of TF-IDF values which were dependent on the source of the document, under the assumption that some words, such as CNN, would be more frequent based on who wrote the document. Beyond modifying the TF-IDF values, the similarity measure was also extended to include some normalization techniques that take into consideration the source of the documents, and the average similarity of documents. Finally, their model was extended

to compare segmentations of documents, rather than entire documents. Overall, these modifications showed noticeable improvements over the basic usage of TF-IDF values and similarity metrics.

Kumaran and Allan expand on Brants et al. by considering not only the term frequencies when computing the similarity between documents, but also named entities, such as “President Obama,” in the documents[11]. Two additional vectors are created for each document: One composed of just the named entities occurring in a document, and another composed of all words that are not named entities. When comparing the similarity between two documents, the standard vector is initially used, and then the similarity between the additional vectors are used to provide finer distinctions, such as whether two documents refer to the same set of named entities, and the same set of general topics, i.e. all the non named entities.

In [12], Lam et al. extend a document-space approach by associating each document with three vectors: a TF-IDF weighted vector; a TF-IDF score of named entities present in the document, similar to [11]; and a concept vector, which details which abstract concepts are contained in the document, using TF-IDF scores based on the frequency of concepts rather than words. The key terms in a document are each given a weight based on which key terms the document contains. Event detection is done by clustering documents as they appear. Each cluster is said to represent a specific event; and documents that do not fit into one cluster are said to be new events. Chen et al. use a similar clustering for event detection but use sentences rather than entire document[2].

Makkonen et al. augment the document representation by using an existing ontology to extract out locations, proper names, and temporal references from the document[15]. These three, combined with the remaining terms in the document are used as the basis for comparison.

Overall, the current event detection systems that do not utilize a semantic space have the key benefit of being able to process documents continuously, since no reduction step is required for vector representations. But the key difference is the focus on comparisons between documents, and not words that occur in documents. These approaches must handle different challenges, such as documents that discuss multiple events and elements of documents that are vague but important for distinguishing events.

## 6 Discussion

The semantic space model we have presented has a number of benefits and drawbacks compared to other semantic space and document based techniques for automatic event detection. The most significant outstanding question is how to analyze all the semantic slices produced in an efficient

Lebanon	nuclear	Wii
Exercising great restraint, they instantly launched airstrikes on Lebanon, damaging critical roads, [...], power stations, etc. Hezbollah retaliated with a stream of rockets that penetrated as far Ashaifa, causing a great deal of terror to Israeli civilians.	Pyongyang would not hold negotiations to resolve the outstanding issues with Washington, [...] "the Americans never recognized our security and we were forced to conduct nuclear test to defend ourselves."	The Wii was the name of the console.. It was time to see if it could deliver its promise of "changing the way we game." Their presentation was probably the most "fun," cutting right to the chase and demonstrating uses of the controller in games.
[T]heir goal is to move public opinion in Lebanon against Hezbollah due to the destruction "they" caused the country, establish a strong deterrence for any future attacks, and , of course, destroy as much of Hezbollah's infrastructure and weapons as possible.	Korea nuclear test hasn't tipped military balance [...] hours ago questions surrounding North Korea and its nascent nuclear weapons program took center stage Monday night	It turns out, according to eyewitness reports from the show floor, that we might not have been playing Wii consoles at the Nintendo booth...should I feel betrayed?

**Table 3:** Blog snippets describing events associated with "Lebanon" in July, "nuclear" in October, and "Wii" in May

manner that exposes events. While our preliminary analysis has shown that events can be detected using TRI, our approach does not currently scale to searching across all terms, nor to identifying events for new words that infrequently occur. However, we argue that the advantages provided by TRI outweigh the outstanding issues and merits further work to address these limitations.

Event detection systems that use semantic spaces have two notable challenges due to how time is integrated. First, the space must be easily modifiable as new documents are produced. Existing approaches use a single dimensionality reduction step after a corpus had been processed to improve information retrieval. However this step limits the integration of new documents into the semantic space; to integrate new documents, the space must be completely recomputed. The second challenge stems from comparing word meanings and documents that occur in different times. Approaches such as [10, 5] that arbitrarily segment the corpora used into different semantic spaces artificially limit both the types of comparisons available and the specific time ranges of the semantics. TRI addresses both of these challenges efficiently. By being based on Random Indexing, dimensionality reduction is done concurrently with developing semantic vectors. Additionally, by utilizing the same set of index vectors over all documents analyzed, every semantic slice is contained within the same semantic space, avoiding the need for reference only those vectors that are common to several time periods.

Conversely, the document based methods discussed in section 5 provided a means of avoiding a post processing stage by incrementally determining the TF-IDF values for words in the corpus. While these approaches efficiently allow the inclusion of more documents over time, each document vector encounters similar problems seen in basic co-occurrence semantic space models, most notably the requirement that two documents have the same exact words for them to be declared similar.

The introduction of additional vector representations of a document, such as the named entity vectors, or the concept vectors, attempt to address this issue, but these additions allude to benefits provided by a semantic space model. For instance, if two documents describe the same events, but without using the same set of words, and instead use highly similar words to describe the event differently, the document based event detection methods would either report two distinct events, or rely on some system which can determine the similarity between two words. Being based on word semantics, TRI avoids this problem, and provides a way of determining how similar two terms are, or which

concepts a word refers to. With TRI, synonymous key words describing the event are modified in a similar manner, and words with similar meanings will have similar effects on the semantics. It may also be possible with TRI to detect synonymous event names by identifying words with similar shifts and similar neighbors. However, further investigation is needed.

While TRI provides elegant solutions to several problems in event detection, significant questions still remain. First, a suitable method of analyzing the semantic shift between vectors is needed. Our initial experiment illustrates tracking outliers based on cosine similarity works well in practice; however, this does not utilize all the information present and could leave some events undetected. Time series analysis or probability distribution analysis are two techniques which might be well suited for similarity comparisons between semantic slices. However, it remains an open question of what limitations exist to the types of events TRI can be detected, and whether the method of comparison can be targeted to find specific kinds of events.

As a second issue, the relationship should be established between the corpus, the duration of a semantic slice, and the types of events that are detected. Our current system was able to detect changes at a monthly granularity, but real-time event detection must operate on a much finer scale. Further work is needed to determine how brief a semantic slice can be while still adequately representing the semantics necessary for event detection.

Regarding the granularity of semantic slices and semantic vectors, we suspect that the optimal granularity is highly dependent on how dense documents are with regards to time in the corpus. One drawback of Random Indexing is the need for a large amount of data, and if there is not enough data, semantic vectors become poorly defined and produce weak similarity scores. We found that the corpus used in the experiment was sparse enough to produce a degradation in the semantics when our semantic slices were set to a time range shorter than a month. Ideally the corpus should have a dense enough set of topics for very narrow semantic slices.

## 7 Conclusion

Unstructured, unfiltered corpora such as blogs present an ideal opportunity for automated event-detection systems to identify new events before they can be reported through more formal sources. We have presented an algorithm that uses changes in word semantics to detect new events

in blog posts. Our approach utilizes simple word co-occurrence and scales well to processing millions of blog posts. Additionally, initial experiments to identify events for specific words proved successful. Further work is needed to identify the strengths and weakness of this approach and quantify its ability to detect events. However, we plan to address these issues in future work. Last we plan to release the implementation of TRI as a part of the S-Space Package[8].

## Acknowledgements

We thank John Cho and his lab for access to the blogline corpus used in this work. We also thank the anonymous reviewers for their comments and suggestions.

## References

- [1] T. Brants, F. Chen, and A. Farahat. A system for new event detection. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 330–337, New York, NY, USA, 2003. ACM.
- [2] K.-Y. Chen, L. Luesukprasert, Seng-cho, and T. Chou. Hot Topic Extraction Based on Timeline Analysis and Multidimensional Sentence Modeling. *IEEE Transactions on Knowledge and Data Engineering*, 19(8):1016–1025, 2007.
- [3] T. Cohen and D. Widdows. Empirical distributional semantics: Methods and biomedical applications. *Journal of Biomedical Informatics*, 42(2):390–405, 2009.
- [4] J. R. Firth. *A synopsis of linguistic theory 1930-1955*. Oxford: Philological Society, 1957. Reprinted in F. R. Palmer (Ed.), (1968). Selected papers of J. R. Firth 1952-1959, London: Longman.
- [5] B. Fortuna, D. Mladenić, , and M. Grobelnik. Visualization of temporal semantic spaces. In J. Davies, M. Grobelnik, and D. Mladenić, editors, *Semantic Knowledge Management*, pages 155–169. Springer Berlin Heidelberg, 2009.
- [6] Z. Harris. *Mathematical Structures of Language*. Wiley, New York, 1968.
- [7] D. B. III and L. N. Trefethen. *Numerical linear algebra*. Philadelphia: Society for Industrial and Applied Mathematics, 1997.
- [8] D. Jurgens and K. Stevens. The S-Space Package: An open source package for semantic spaces. <http://code.google.com/p/airhead-research/>.
- [9] P. Kanerva, J. Kristoferson, and A. Holst. Random indexing of text samples for latent semantic analysis. In L. R. Gleitman and A. K. Josh, editors, *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, page 1036, 2000.
- [10] A. Kontostathis, I. De, L. E. Holzman, and W. M. Pottinger. Use of term clusters for emerging trend detection. Technical report, Lehigh University, 2004.
- [11] G. Kumaran and J. Allan. Text classification and named entities for new event detection. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 297–304. ACM, 2004.
- [12] W. Lam, H. Meng, K. L. Wong, and J. Yen. Using contextual analysis for news event detection. *International Journal on Intelligent Systems*, 16(44):525–546, 2001.
- [13] T. K. Landauer and S. T. Dumais. A solution to Plato’s problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240, 1997.
- [14] M. L. Littman, S. T. Dumais, and T. K. Landauer. Automatic cross-language information retrieval using latent semantic indexing. In G. Grefenstette, editor, *Cross language information retrieval*. Kluwer, 1998.
- [15] J. Makkonen, H. Ahonen-Myka, and M. Salmenkivi. Simple Semantics in Topic Detection and Tracking. *Information Retrieval*, 7:347–368, 2004.
- [16] D. L. T. Rohde, L. M. Gonnerman, and D. C. Plaut. An improved model of semantic similarity based on lexical co-occurrence. *Cognitive Science*, 2009. submitted.
- [17] E. Sagi, S. Kaufmann, and B. Clark. Semantic density analysis: Comparing word meaning across time and phonetic space. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 104–111, Athens, Greece, March 2009. Association for Computational Linguistics.
- [18] M. Sahlgren. Vector-based semantic analysis: Representing word meanings based on random labels. In *Proceedings of the ESSLLI 2001 Workshop on Semantic Knowledge Acquisition and Categorisation*, Helsinki, Finland, 2001.
- [19] M. Sahlgren, A. Holst, and P. Kanerva. Permutations as a means to encode order in word space. In *Proceedings of the 30th Annual Meeting of the Cognitive Science Society (CogSci'08)*, 2008.
- [20] M. Sahlgren and J. Karlgren. Buzz Monitoring in Word Space. In *Proceedings of European Conference on Intelligence and Security Informatics (EuroISI 2008)*, Esbjerg, Denmark, 2008.
- [21] H. Schütze and J. O. Pedersen. A cooccurrence-based thesaurus and two applications to information retrieval. *Information Processing and Management*, 33(3):307–318, 1997.
- [22] K. C. Sia, J. Cho, Y. Chi, and B. L. Tseng. Efficient computation of personal aggregate queries on blogs. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, August 2008.
- [23] A. Vinokourov, J. Shawe-Taylor, and N. Cristianini. Finding Language-Independent Semantic Representation of Text using Kernel Canonical Correlation Analysis. Technical report, Neurocolt, 2002. NeuroCOLT Technical Report NC-TR-02-119.



# A Pairwise Event Coreference Model, Feature Impact and Evaluation for Event Coreference Resolution

Zheng Chen  
The Graduate Center  
The City University of New York  
zchen1@gc.cuny.edu

Heng Ji  
Queens College and The Graduate Center  
The City University of New York  
hengji@cs.qc.cuny.edu

Robert Haralick  
The Graduate Center  
The City University of New York  
haralick@aim.com

## Abstract

In past years, there has been substantial work on the problem of entity coreference resolution whereas much less attention has been paid to event coreference resolution. Starting with some motivating examples, we formally state the problem of event coreference resolution in the ACE<sup>1</sup> program, present an agglomerative clustering algorithm for the task, explore the feature impact in the event coreference model and compare three evaluation metrics that were previously adopted in entity coreference resolution: MUC F-Measure, B-Cubed F-Measure and ECM F-Measure.

## Keywords

Pairwise Event Coreference Model, Event Coreference Resolution, Event Attribute

## 1. Introduction

In this paper, we address the task of event coreference resolution specified in the Automatic Content Extraction (ACE) program: grouping all the mentions of events in a document into equivalent classes so that all the mentions in a given class refer to a unified event. We adopt the following terminologies used in ACE [1]:

- Entity: an object or set of objects in the world, such as person, organization, facility.
- Event: a specific occurrence involving participants.
- Event trigger: the word that most clearly expresses an event's occurrence.
- Event argument: an *entity*, or a *temporal expression* or a *value* that has a certain role (e.g., PLACE) in an event.
- Event mention: a sentence or phrase that mentions an event, including a distinguished trigger and involving arguments. An event is a cluster of event mentions.
- Event attributes: an event has six event attributes, event type, subtype, polarity, modality, genericity, and tense.

We demonstrate some motivating examples in table 1 (event triggers are surrounded by curly brackets and event arguments are underlined).

In example 1, event mention *EM1* corefers with *EM2* because they have the same event type and subtype

(CONFLICT: ATTACK) indicated by two verb triggers “tore” and “exploded” respectively, and the argument “a waiting shed” in *EM1* corefers with “the waiting shed” in *EM2*. In example 2, *EM1*, *EM2* and *EM3* corefer with each other because they have the same event type and subtype (LIFE:MARRY) indicated by a verb trigger “wed” and two noun triggers “ceremony” and “nuptials” respectively. Furthermore, the two persons “Rudolph Giuliani” and “Judith Nathan” involving in the “Marry” event in *EM1* corefer with “Giuliani” and “Nathan” in *EM3* respectively. In example 3, *EM1* does not corefer with *EM2* although they have the same event type and subtype (LIFE:DIE) because the event attribute “polarity” of *EM1* is “POSITIVE” (occurred) while in *EM2*, it is “NEGATIVE” (not occurred).

Table 1. Motivating examples for event coreference resolution

<p><i>Example1</i></p> <p><i>EM1</i>: <u>A powerful bomb</u> {tore} through <u>a waiting shed</u> at <u>the Davao City international airport</u>.</p> <p><i>EM2</i>: <u>The waiting shed</u> literally {exploded}.</p>
<p><i>Example2</i></p> <p><i>EM1</i>: <u>Rudolph Giuliani</u> will {wed} his companion, <u>Judith Nathan</u>, on <u>May 24</u> in <u>the ex-mayor's old home</u>.</p> <p><i>EM2</i>: Mayor Michael Bloomberg, will perform the {ceremony}.</p> <p><i>EM3</i>: The <u>Giuliani-Nathan</u> {nuptials} will be a first for Bloomberg, who is making an exception from his policy of not performing weddings.</p>
<p><i>Example3</i></p> <p><i>EM1</i>: <u>At least 19 people</u> were {killed} in the first blast.</p> <p><i>EM2</i>: There were no reports of {deaths} in the second blast.</p>

The major contributions of this paper are:

- (1) A formal statement of event coreference resolution and an algorithm for the task.
- (2) A close study of four event attributes: polarity, modality, genericity and tense.
- (3) A close study of feature impact on the performance of the pairwise event coreference model.

<sup>1</sup> <http://www.nist.gov/speech/tests/ace/>

## 2. Event Coreference Resolution

We formulate the problem of event coreference resolution as an agglomerative clustering task. The basic idea is to start with singleton event mentions, traverse through each event mention (from left to right) and iteratively merge the active event mention into a prior established event or start the event mention as a new event. We first introduce the notation needed for our algorithm.

### 2.1 Notation

Let  $I$  be the set of positive integers. Let  $A$  be a set of attributes and  $V$  be a set of values. Some attributes may have no values and some attributes may have one or more values. Any information about an event is a subset of  $A \times V$ , and the same applies to an event mention. Such abstraction makes it possible for us to extend the meaning of attributes.

In this paper we state that an event includes the following attribute members: 6 event attributes (type, subtype, modality, polarity, genericity and tense), a set of arguments, and a set of event mentions. Accordingly, we have the following notation:

Let  $e$  be an ACE event. Let  $e.arg$  be the set of arguments (*entities, temporal expressions and values*) in the event  $e$ . Let  $e.ems$  be the set of event mentions in the event  $e$ .

An event mention has a distinguished trigger and a set of arguments. Accordingly, we have the following notation:

Let  $em$  be an event mention. Let  $em.trigger$  be the event trigger. Let  $em.arg$  be the set of arguments (*entities, temporal expressions and values*) in the event mention  $em$ .

Let  $M$  be the set of possible event mentions in a document  $D$ . Let  $\langle em_i \in M \mid i = 1, \dots, N \rangle$  be the  $N$  event mentions in the document  $D$  listed in the order in which they occur in the document.

Let  $E$  be the set of possible events in the document  $D$ . Let  $\langle e_j \in E \mid j = 1, \dots, K \rangle$  be the  $K$  events.

The goal of event coreference resolution is to construct a function  $f: I \rightarrow I$ , mapping event mention index  $i \in I$  to event index  $j \in I$ .

Initially, each event mention  $em$  is wrapped in an event  $e'$  so that  $e'$  contains a single event mention  $em$ . We denote the wrapping function as  $\alpha: M \rightarrow E'$ , i.e.,  $e' = \alpha(em)$  where  $E'$  is the set of  $e'$ . Furthermore,  $em$  and  $e'$  satisfy the following properties: (1)  $e'.ems = em$  (2)  $e'.arg = em.arg$

### 2.2 Algorithm

We describe an agglomerative clustering algorithm that gradually builds up the set of events by scanning each event mention from left to right.

Let  $E_0$  be the initial set of established events and  $E_0 = \emptyset$ .  $E_1 = \{\alpha(em_1)\}$  and  $f(1) = 1$ . Let  $\delta$  be a threshold. Let

$coref: E \times M \rightarrow (0,1)$  be a function which gives a score to any (event, event mention) pair.

At each iteration  $k$  ( $k = 2, \dots, N$ ), let  $e_j \in E_{k-1}$  satisfy

$coref(e_j, em_k) \geq coref(e, em_k)$  for any  $e \in E_{k-1}$

If  $coref(e_j, em_k) \geq \delta$ , then  $f(k) = j$  and

$$E_k = \{e_1^k, \dots, e_{N_k}^k\}$$

where  $e_n^k = e_n^{k-1}$  for  $n \neq j$  and  $e_n^k.ems = e_n^{k-1}.ems \cup \{em_k\}$ ,  $e_n^k.arg = e_n^{k-1}.arg \cup em_k.arg$  for  $n = j$ .

If  $coref(e_j, em_k) < \delta$ , then  $f(k) = N_{k-1} + 1$  and

$$E_k = E_{k-1} \cup \{e_{N_k}^k\}$$

where  $N_k = N_{k-1} + 1$  and  $e_{N_k}^k = \alpha(em_k)$

After  $N - 1$  iterations, we resolve all the event coreferences in the document.

The complexity of the algorithm is  $O(N^2)$ . However, if we only consider those event mentions with the same event type and subtype, we can decrease its running time.

### 2.3 Pairwise Event Coreference Model

A key issue in the above algorithm is how to compute the coreference function  $coref(\cdot, \cdot)$  which indicates the coreference score between the active event mention and a prior established event. We construct a Maximum-entropy model for learning such function. The features applied in our model are tabulated in Table 2. We categorize our features into *base, distance, arguments* and *attributes* feature sets to capture trigger relatedness, trigger distance, argument compatibility and event attribute compatibility respectively.

In this paper, we run NYU's 2005 ACE system [2] to tag event mentions. However, their system can only extract triggers, arguments and two event attributes (event type and subtype) and cannot extract the other four event attributes. Therefore, we developed individual components for those four event attributes (polarity, modality, genericity and tense). Such efforts have been largely neglected in the prior research due to their low weights in the ACE scoring metric [1]. The event attributes absolutely play an important role in event coreference resolution because two event mentions cannot corefer with each other if any of the attributes conflict with each other. We encode the event attributes as features in our model and study their impact on the system performance. In the next section, we describe the four event attributes in details.

## 3. Extracting the Four Event Attributes

### 3.1 Polarity

An event is NEGATIVE if it is explicitly indicated that the event did not occur, otherwise, the event is POSITIVE. The following list reviews some common ways in which NEGATIVE polarity may be expressed (triggers are

**Table 2. Feature categories for the pairwise event coreference model**

Category	Features	Feature Values ( <i>aem</i> : the active event mention, <i>e</i> : a partially-established event, <i>lem</i> : the last event mention in <i>e</i> )
Base	type_subtype	pair of event type and subtype in <i>aem</i>
	nominal	1 if the trigger of <i>aem</i> is nominal
	nom_number	plural or singular if the trigger of <i>aem</i> is nominal
	pronominal	1 if the trigger of <i>aem</i> is pronominal
	exact_match	1 if the trigger spelling of <i>aem</i> matches the trigger spelling of an event mention in <i>e</i>
	stem_match	1 if the trigger stem in <i>aem</i> matches the trigger stem of an event mention in <i>e</i>
	trigger_sim	the maximum of quantized semantic similarity scores (0-5) using WordNet resource among the trigger pairs of <i>aem</i> and an event mention in <i>e</i>
	trigger_pair	trigger pair of <i>aem</i> and <i>lem</i>
	pos_pair	part-of-speech pair of triggers of <i>aem</i> and <i>lem</i>
Distance	token_dist	how many tokens between triggers of <i>aem</i> and <i>lem</i> (quantized)
	sentence_dist	how many sentences <i>aem</i> and <i>lem</i> are apart (quantized)
	event_dist	how many events in between <i>aem</i> and <i>lem</i> (quantized)
Arguments	overlap_num, overlap_roles	overlap number of arguments and their roles (role and id exactly match) between <i>aem</i> and <i>e</i>
	prior_num, prior_roles	the number of arguments that only appear in <i>e</i> and their roles
	act_num, act_roles	the number of arguments that only appear in <i>aem</i> and their roles
	coref_num	the number of arguments that corefer with each other but have different roles between <i>aem</i> and <i>e</i>
	time_conflict	1 if both <i>aem</i> and <i>e</i> have an argument with role “Time-Within” and their values conflict
	place_conflict	1 if both <i>aem</i> and <i>e</i> have an argument with role “Place” and their values conflict
Attributes	mod,pol,gen, ten	four event attributes in <i>aem</i> : modality, polarity, genericity, and tense
	mod_conflict, pol_conflict, gen_conflict, ten_conflict	four boolean values indicating whether the attributes of <i>aem</i> and <i>e</i> conflict

surrounded by curly brackets, the words indicating NEGATIVE are underscored)

- Using a negative word such as not, no

*Guns don't {kill} people, people do.*

*No death sentence has ever been {executed} in the country.*

- Using context, e.g., the embedding predicate with a negative meaning or sentence patterns

*Bush indefinitely postponed a {visit} to Canada.*

*She had decided to stay home rather than {go} to a dance.*

### 3.2 Modality

An event is ASSERTED if it is mentioned as if it were a real occurrence, otherwise it is OTHER. Two “ASSERTED” examples are listed as follows:

*At least 19 people were {killed} in Tuesday's blast.*

*We condemn all {attacks} against civilians in Haifa.*

The “OTHER” examples have much more varieties. The examples include, but are not limited to (triggers are surrounded by curly brackets, the words indicating modality are underscored)

- believed events

*I believe he will be {sentenced}.*

- hypothetical events

*If convicted of the killings, Vang {faces} life in prison.*

- commanded and requested events

*He was commanded to {leave} his country.*

- threatened, proposed and discussed events

*He was threatened to {pay} the ransom.*

- desired events

*He desires to be {elected}.*

- promised events

*The terrorist said he would {attack} the village.*

The modality of events can be characterized by a veridicality axis that ranges from truly factual to counterfactual and a spectrum of modal types fall between the two extremes, expressing *degrees of possibility, belief, evidentiality, expectation, attempting, and command* [3]. Actually, ACE has largely simplified the problem, i.e., the modality is “ASSERTED” for the two extremes, and is “OTHER” for all the other modal types.

### 3.3 Genericity

An event is SPECIFIC if it is a single occurrence at a particular place and time, or a finite set of such occurrences; otherwise, it is GENERIC.

Some GENERIC examples are listed as follows:

*Hamas vowed to continue its {attacks}.*

*Roh has said any pre-emptive {strike} against the North's nuclear facilities could prove disastrous.*

### 3.4 Tense

The tense of events can be characterized by a temporal axis in which we define the time of publication or broadcast as the *textual anchor time*. The PAST events occurred prior to the anchor time; the FUTURE events have not yet occurred at the anchor time; the PRESENT events occur at the anchor time; all the other events are UNSPECIFIED.

### 3.5 Models for the Four Event Attributes

We construct a Maximum-entropy model for each of the four event attributes. All the models apply the following common features:

- the trigger and its part-of-speech
- event type and subtype
- the left two words of the trigger (lower case) and their POS tags
- the right two words of the trigger (lower case) and their POS tags

Furthermore, the polarity model also applies the following two features:

- the embedding verb of the trigger if any
- a boolean feature indicating whether a negative word exists (not, no, cannot or a word ending with n't) ahead of the trigger and within the clause containing the trigger.

The modality model also applies the following feature:

- a boolean feature indicating whether a modal auxiliary (may, can, etc.) or modal adverbs (possibly, certainly, etc.) exists ahead of the trigger and within the clause containing the trigger.

The genericity model also applies the following three features:

- a boolean feature indicating whether the event mention has a “PLACE” argument
- a boolean feature indicating whether the event mention has a “TIME-WITHIN” argument
- the number of arguments that the event mention has except “PLACE” and “TIME-WITHIN”

The tense model also applies the following two features:

- the first verb within the clause containing the trigger and its POS tag
- the head words of the “TIME-WITHIN” argument if the event mention has one

## 4. Experiments and Results

### 4.1 Data and Evaluation Metrics

For our experiments, we used the ACE 2005 English corpus which contains 599 documents in six genres: newswire, broadcast news, broadcast conversations, weblogs, newsgroups and conversational telephone speech transcripts. We first investigated the performance of the four event attribute classification models using the ground truth event mentions and system generated event mentions respectively. The evaluation metrics we adopted in this set of experiments are Precision (P), Recall (R) and F-Measure (F). We then validated our agglomerative clustering algorithm for the event coreference resolution using the ground truth event mentions and system generated event mentions respectively. The evaluation metrics we adopted in this set of experiments are three conventional metrics for entity coreference resolution, namely, MUC F-Measure [4], B-Cubed F-Measure [5] and ECM F-Measure [6]. We conducted all the experiments by ten times ten-fold cross validation and measured significance with the Wilcoxon signed rank test.

### 4.2 Performance of the Four Event Attribute Classification Models

Table 3 shows that the majority of event mentions are POSITIVE (5162/5349=0.965), ASSERTED (4002/5349=0.748), SPECIFIC (4145/5349=0.775) and PAST (2720/5349=0.509).

**Table 3. Statistics of the four event attributes in the corpus**

Attribute	Instance counts in the ACE corpus
Polarity	NEGATIVE=187, POSITIVE=5162
Modality	ASSERTED=4002, OTHER=1347
Genericity	GENERIC=1204, SPECIFIC=4145
Tense	FUTURE=593, PAST=2720, PRESENT=152, UNSPECIFIED=1884

Table 4 shows the performance of the four event attribute classification models using the ground truth event mentions (perfect) and the system generated event mentions (system). For comparison, we also set up a



**Table 4. Performance of the four event attribute classification models**

	Polarity			Modality			Genericity			Tense		
	P	R	F	P	R	F	P	R	F	P	R	F
Perfect (majority)	0.966	1.0	0.983	0.748	1.0	0.856	0.777	1.0	0.874	0.510	1.0	0.675
Perfect (model)	0.968	1.0	0.984	0.784	1.0	0.879	0.795	1.0	0.885	0.644	1.0	0.783
System (majority)	0.969	0.573	0.720	0.779	0.519	0.622	0.792	0.523	0.629	0.550	0.432	0.483
System (model)	0.974	0.574	0.722	0.805	0.527	0.637	0.799	0.525	0.633	0.677	0.484	0.564

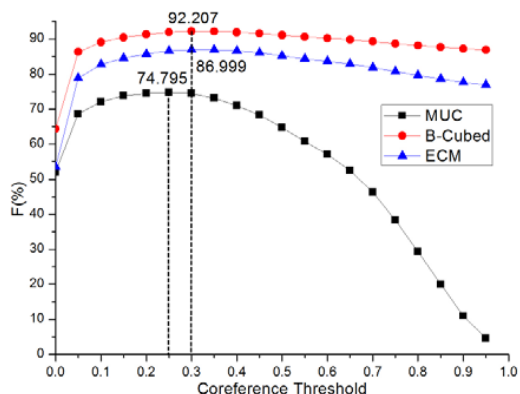
baseline for each case using the majority value as output (e.g., for Polarity attribute, we always set the value to POSITIVE because POSITIVE is the majority).

Table 4 shows that the improvements for Polarity, Modality and Genericity over the baselines are quite limited while the improvements for Tense are significant, either using ground truth event mentions or using system generated event mentions.

### 4.3 Determining Coreference Threshold $\delta$

In order to determine the best coreference threshold  $\delta$  in our agglomerative clustering algorithm, we conducted this set of experiments by integrating full feature sets (as listed in Table 2) in the pairwise event coreference model. We investigate how the performance varies by adjusting the coreference threshold  $\delta$ . For this set of experiments, we use ground truth event mentions.

Figure 1 shows the F-scores based on the three evaluation metrics by varying the coreference threshold  $\delta$ . The best MUC F-score, B-Cubed F-score and ECM F-score are obtained at  $\delta = 0.25$ ,  $\delta = 0.3$ ,  $\delta = 0.3$  respectively. It is worth noting that the MUC F-score drops dramatically after  $\delta = 0.5$ . We observed that as the threshold increases, more singleton events are produced and the dramatic decrease in MUC recall cannot offset the increase in MUC precision. As [5], [6] have pointed out, MUC metric does not give any credit for separating out singletons, therefore it is not quite effective in evaluating system responses with many singletons. The B-Cubed curve shows similar fluctuations compared to the ECM curve.



**Figure 1. Determining the best coreference threshold  $\delta$**

### 4.4 Feature Impact

Table 5 presents the impact of aggregating feature sets on the performance of our pairwise event coreference model using the ground truth event mentions (coreference threshold  $\delta = 0.3$ ).

**Table 5. Feature impact using ground truth event mentions**

	MUC F	B-Cubed F	ECM F
Base	0.386	0.868	0.777
+Distance	0.446	0.866	0.781
+Arguments	0.530	0.879	0.804
+Attributes	0.723	0.919	0.865

Our Wilcoxon signed rank tests show that the F-score improvements are significant for all three metrics when we apply richer features except that there is a little deterioration for the distance feature set using B-Cubed metric. We observe that the improvement is dramatic using the MUC metric. However, it is not quite reasonable since we evaluate on the same system responses, varying in metrics. Since ECM overcomes some shortcomings of MUC and B-Cubed metrics as explained in [6], we focus on analyzing the results from ECM metric. In this setting, distance feature set contributes about 0.4% F-score improvement, while arguments feature set contributes nearly 2.4% F-score improvement. It is clear that the attribute feature set contributes the most significant contribution (6.08% absolute improvement).

We then investigate whether the feature sets have similar impacts on the pairwise event coreference model using the system generated event mentions.

**Table 6. Feature impact using system generated event mentions**

	MUC F	B-Cubed F	ECM F
Base	0.265	0.558	0.489
+Distance	0.254	0.548	0.483
+Arguments	0.274	0.552	0.490
+Attributes	0.28	0.554	0.492

Table 6 shows that the aggregated features do not bring great improvements using the system generated event mentions. The reason is that the spurious and missing event mentions labeled by the event extractor not only

directly affect the final score of event coreference, but also lead to the deteriorated event coreference model which is learned from spurious feature values. We name the spurious event coreference caused by the spurious event mentions as *type I error*, the spurious event coreference caused by the model and the clustering algorithm as *type II error*. Similarly, we define *type I miss* and *type II miss*, one is caused by the missing event mentions and the other is caused by the model and the algorithm. Table 7 shows the average ratio of *type I error* and *type II error*, ratio of *type I miss* and *type II miss* for each model which only applies the features in each feature category. It is clear that the performance bottleneck of event coreference resolution comes from the performance of system event mentions.

**Table 7. Ratio of type I,II error and ratio of type I,II miss**

	<i>type I error</i> Vs. <i>type II error</i>	<i>type I miss</i> Vs. <i>type II miss</i>
Base	90%/10%	82.3%/17.7%
Distance	99.8%/0.2%	77.5%/22.5%
Arguments	95.2%/4.8%	81%/19%
Attributes	92.3%/7.7%	79.6%/20.4%

## 5. Related Work

Earlier work on event coreference (e.g. [7], [8]) in MUC was limited to several scenarios, e.g., terrorist attacks, management succession, resignation. The ACE program takes a further step towards processing more fine-grained events. Ahn presented an event extraction system in which event coreference resolver is located at the end of a pipeline of event extraction [9]. However, Ahn did not point out what evaluation metric he used. [10] presented a graph-based method for event coreference resolution and proposed two methods for computing the coreference score between two event mentions. However, they only reported evaluation results on ground-truth event mentions.

Our experiments show that a high-performance event coreference resolver relies on a high-performance event mention extractor. Earlier work on event extraction systems was presented in [9],[11],[12],[13],[14].

## 6. Conclusions and Future Work

We have formally stated the problem of event coreference resolution, presented an algorithm involving a pairwise event coreference model and studied the feature impacts on the pairwise event coreference model.

In the future, we will continue to put great efforts on improving the performance of event extraction system including trigger labelling, argument labelling and event attribute labelling. We believe that the improved components will finally help us improve the performance of event coreference resolution.

## 7. Acknowledgements

This material is based upon work supported by the Defense

Advanced Research Projects Agency under Contract No. HR0011-06-C-0023 via 27-001022, Google Research, CUNY Research Enhancement Program and GRTI Program. Any opinions, findings and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the U. S. Government.

## 8. References

- [1] NIST. 2005. The ACE 2005 Evaluation Plan. <http://www.itl.nist.gov/iad/mig/tests/ace/ace05/doc/ace05-evalplan.v3.pdf>.
- [2] R. Grishman, D. Westbrook, and A. Meyers. 2005. NYU's English ACE 2005 System Description. In *ACE 05 Evaluation Workshop*, Gaithersburg, MD.
- [3] R. Sauri, M. Verhagen and J. Pustejovsky. 2006. Annotating and Recognizing Event Modality in Text. In *Proceedings of the 19<sup>th</sup> International FLAIRS Conference, FLAIRS 2006*. Melbourne Beach, Florida. May 11-13, 2006.
- [4] M. Vilain, J. Burger, J. Aberdeen, D. Connolly and L. Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*.
- [5] A. Bagga and B. Baldwin. 1998. Algorithms for scoring coreference chains. *Proc. The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*.
- [6] X. Luo. 2005. On coreference resolution performance metrics. *Proc. of HLT-EMNLP*.
- [7] A. Bagga and B. Baldwin. 1999. Cross-document event coreference: Annotations, experiments, and observations. In *Proc. ACL-99 Workshop on Coreference and Its Applications*.
- [8] K. Humphreys, R. Gaizauskas, S. Azzam. 1997. Event coreference for information extraction. In *Proceedings of the ACL Workshop on Operational Factors in Practical Robust Anaphora Resolution for Unrestricted Texts*.
- [9] D. Ahn. 2006. The stages of event extraction. *Proc. COLING/ACL 2006 Workshop on Annotating and Reasoning about Time and Events*. Sydney, Australia.
- [10] Z. Chen and H. Ji. 2009. Graph-based Event Coreference Resolution. *Proc. ACL-IJCNLP 2009 workshop on TextGraphs-4: Graph-based Methods for Natural Language Processing*.
- [11] R. Grishman, D. Westbrook and A. Meyers. 2005. NYU's English ACE 2005 System Description. *Proc. ACE 2005 Evaluation Workshop*. Washington, US.
- [12] H. Ji and R. Grishman. 2008. Refining Event Extraction Through Cross-document Inference. *Proc. ACL 2008*. Ohio, USA.
- [13] Z. Chen and H. Ji. 2009. Language Specific Issue and Feature Exploration in Chinese Event Extraction. *Proc. HLT-NAACL 2009*. Boulder, Co.
- [14] Z. Chen and H. Ji. 2009. Can One Language Bootstrap the Other: A Case Study on Event Extraction. *Proc. HLT-NAACL Workshop on Semi-supervised Learning for Natural Language Processing*. Boulder, Co.

# Summarizing Threads in Blogs Using Opinion Polarity

Alexandra Balahur<sup>1,2</sup>, Elena Lloret<sup>1</sup>, Ester Boldrini<sup>1</sup>,  
Andrés Montoyo<sup>1</sup>, Manuel Palomar<sup>1</sup>, Patricio Martínez-Barco<sup>1</sup>

<sup>1</sup>Natural Language Processing and Information Systems Group  
Dept. of Software and Computing Systems, University of Alicante  
Apartado de Correos 99, E-03080, Alicante, Spain

<sup>2</sup>European Commission Joint Research Centre  
Institute for the Protection and Security of the Citizen  
Global Security and Crisis Management Unit, OPTIMA Action  
Via E. Fermi, 2749, I-21027, Ispra (VA), Italy  
{abalahur, elloret, eboldrini, montoyo, mpalomar, patricio}@dlsi.ua.es

## Abstract

The huge amount of data available on the Web needs to be organized in order to be accessible to users in real time. This paper presents a method for summarizing subjective texts based on the strength of the opinion expressed in them. We used a corpus of blog posts and their corresponding comments (blog threads) in English, structured around five topics and we divided them according to their polarity and subsequently summarized. Despite the difficulties of real Web data, the results obtained are encouraging; an average of 79% of the summaries is considered to be comprehensible. Our work allows the user to obtain a summary of the most relevant opinions contained in the blog. This allows them to save time and be able to look for information easily, allowing more effective searches on the Web.

## Keywords

Opinion Mining, Sentiment Analysis, Blog Posts, Automatic Summarization.

## 1. Introduction

Due to the rapid development of the Social Web, new textual genres expressing subjective content by means of emotions, feelings, sentiments, moods or opinions are growing rapidly. Nowadays, people converse frequently using many non-conventional ways of communication such as blogs, forums or reviews. As a consequence, the number of such emerging text types is growing at an exponential rate, as well as their impact on the everyday lives on millions of people.

A research for the Pew Institute [1] shows that 75,000 blogs are created per day by people all over the world, on a great variety of subjects. Thus, blogs are becoming an extremely relevant resource for different kinds of studies focused on many useful applications. This research area have become known as *sentiment analysis* or *opinion mining*. However, as there is no overall accepted definition of this task and in order to delimit our research area, the

concepts of emotions, feelings, sentiments, moods and opinions need to be defined with precision.

Emotion is “an episode of interrelated, synchronized changes in the states of all or most of the five organismic subsystems) in response to the evaluation of an external or internal stimulus event as relevant to major concerns of the organism [2, 3].

The term “feeling” points to a single component denoting the subjective experience process [4] and is therefore only a small part of an emotion.

“Moods” are less specific and intense affective phenomena, product of two dimensions - energy and tension [5].

“Sentiment” is defined in the Webster dictionary<sup>1</sup> as: 1 a: an attitude, thought, or judgment prompted by feeling: predilection b: a specific view or notion: opinion; 2 a: emotion b: refined feeling: delicate sensibility especially as expressed in a work of art c: emotional idealism d: a romantic or nostalgic feeling verging on sentimentality; 3 a: an idea colored by emotion b: the emotional significance of a passage or expression as distinguished from its verbal context. Finally, the term “opinion”, according to the Webster Dictionary, is 1 a: a view, judgment, or appraisal formed in the mind about a particular matter b: approval, esteem; 2 a: belief stronger than impression and less strong than positive knowledge b: a generally held view; 3 a: a formal expression of judgment or advice by an expert b: the formal expression (as by a judge, court, or referee) of the legal reasons and principles upon which a legal decision is based.

As we can deduce from these definitions, affect-related concepts are similar in nature and in many cases overlap;

---

<sup>1</sup> <http://www.merriam-webster.com/>

however, we can say that emotion is the super category that includes all other abovementioned concepts.

Language employed in blogs is highly heterogeneous [6]. People with different social backgrounds write them and as a consequence, they contain highly variable and unpredictable language [7]. Moreover, surveys show that they are not entirely written using an informal style; it is only employed for a small part of them and many users aim instead at a more refined style. As we can deduce, these texts offer an example of genuine and spontaneous Natural Language, providing the opportunity of challenging studies focused on solving the problems of its understanding and generation. Not less important to mention is also the fact that blogs contain frequent “copy-pastes” from news sources that are introduced to support a point of view or argument.

It is worth mentioning that those emerging texts are extremely relevant also because bloggers write whatever is on their mind about a wide range of topics [8]. In most of the cases, they aim to share their feelings about an episode of their lives, a “hot” news topic or a product, for example [9]; consequently, these corpora provide an excellent platform research on informal communications [10].

Researchers could exploit this huge amount of data for an enormous number of applications useful for companies, economic institutions, educational centers, politic parties, etc. Companies could use them to discover the customers’ preferences, complaints or to monitor opinions about competitors. Economic institutions could take advantage of this information to predict and control people’s attitude towards relevant economic events, as for example the present economic crisis. Furthermore, educational institutions could employ them to know and understand students’ opinion about teachers, methods or didactic materials, for example. And last but not least, politic institutions or parties would use them to know people’s opinion about laws, bills or to foresee elections results. On the one hand, the growing volume of subjective information available on the Web allows for better and more informed decisions of the users, but on the other hand, the quantity of data to be analyzed imposes the automation of the opinion mining process as well as other Natural Language Processing (NLP) tasks. Our research is focused on opinion summarization of blog posts about different topics. Our main purpose is to provide the user with a summary of positive and negative opinions about a specific topic. The summary will be generated in three sizes, 10%, 15% and 20%. Depending on the user profile and its needs we would offer him the size s/he needs. For example, if we work with a blog about mobile phones we would give back a short summary if the user does not have a high level of knowledge of this product; but if the user is a technician, the system would give him back a more detailed summary, because s/he would be able to

understand a more technical and detailed summary. In general, this would avoid spending much of their time reading all the reviews to find what they are looking for, as the system offers them summaries of pros and cons of a topic. This would be one of the possible ways to exploit the huge amount of data the Web offers.

## 2. Motivation and contribution

The explosive increase in Web communication has attracted interest in technologies for automatically mining personal opinions from different kinds of Web documents, such as product reviews, blogs or forums. These technologies would benefit users who seek reviews on certain consumer products [11].

In fact, at the time of taking a decision, more and more people search for subjective information expressed on the Web on their matter of interest and base their final decision on the information found [12]. Not less important is also the fact that people interested in news and how they are reflected in the world wide opinion often use both newspaper sources, as well as blogs, in order to follow the development of news and the corresponding opinion. For this reason, we believe opinion summarization could represent a useful tool, on the one hand to help users to take decisions quickly and, on the other hand, this would also be effective to manage the huge amount of data we have.

The first contribution this paper brings is the annotation of a collection of a corpus of blog posts together with the comments given on them (threads) in English about different topics, at the level of opinion, polarity and post/comment, as well as sentence importance. We decided to select five macrotopics that are economy, science and technology, cooking, society, and sport. We obtained a total of 51 documents containing the discussion threads (original posts and the comments made on them). The average number of comments on the post is 33.

After having collected the corpus, we employed a partial version of EmotiBlog, an annotation scheme for emotion detection in non traditional textual genres [13], labeling the opinions of the different users. We decided to employ a partial version of the model to avoid noise. In fact, EmotiBlog is a fine grained model, but for the first step of our research we only need some of the elements of the traditional annotation scheme. Subsequently, we automatically classified the polarity at a sentence and also at a document level and furthermore, we proposed a method to summarize similar opinions grouped for topics. The result is a summary of positive and negative opinions, divided according to their corresponding polarity.



### 3. Related work

The increasing amount of data on the Web needs to be processed in order to help users who are looking for specific information. Therefore, summarization systems are becoming more and more useful because they provide shorter versions of texts, avoiding users wasting their time. Moreover, subjective information has a high presence on the Internet, by means of forums or blogs, among others. A recent application for summarization is to combine this task with Opinion Mining, in order to produce summaries of opinions on a specific topic. Regarding opinion-oriented summaries, subjective linguistic elements have to be detected and classified first, according to their polarity, and then, they have to be grouped in a coherent fragment of text in order to produce the final summary.

Opinion summarization systems that participated in the Text Analysis Conference<sup>2</sup> (TAC) in 2008 such as [14], [15], [16], or [17] followed these steps. However, out of the scope of the TAC competition, we can find other interesting approaches, as well. For instance, in [18] Machine Learning algorithms are used to determine which sentences should belong to a summary, after identifying possible opinion text spans. The useful features to locate opinion quotations within a text included location within the paragraph and document, and the type of words they contained. Similarly, in [19] the relevant features and opinion words with their polarity (whether a positive or a negative sentiment) are identified, and then, after detecting all valid feature-opinion pairs, a summary is produced, but focusing only in movie reviews. Normally, online reviews also contain numerical ratings that users insert when providing their personal opinions about a product or service. In [20] a Multi-Aspect Sentiment model is proposed. This statistical model uses aspect ratings to discover the corresponding topics and extract fragments of text.

Our work differs from the ones abovementioned since we take into account the posts written in real blogs, to further build a summary of the most relevant opinions contained in them, based on their polarity.

### 4. Corpus collection and labeling

The corpus we employed in this study is a collection of 51 blogs extracted from the Web. This is a limited dataset which allows for a preliminary study in the field; however, in our future work we would like to extend it in order to carry out a more in depth research. The blog posts are written in English and have the same structure. Generally, blogs have the following organization: the authors create an initial post containing a piece of news and their opinion on it and subsequently, bloggers reply expressing their

opinions about the topic. In most of the cases, commenting posts are the most subjective texts even if also in its first intervention the author can express its point of view. They can also contain multimodal information, but we decided to take into account only the text; however, the multimodal information analysis could be an interesting research for future work. In our blog corpus annotation, we indicated the *url* from which the thread was extracted it, we then included the initial annotated piece of news and the labeled user comments.

People use this new textual genre to express opinions on a wide range of topics. However, in order to delimitate our work, we were forced to select only few of them; we gave priority to the most relevant threads, that contained a large amount of posts in order to have a considerable amount of data. We chose some of the topics that we considered relevant: economy, science and technology, cooking, society and sport. Regarding its size, Table 1 shows the average and the total number of posts, of words in the news, of the number of words in posts and, finally, of words both in news and in posts.

Table 1: Corpus size

	N. Posts	N. Words for new	N. Word for post	Total words
<b>Total</b>	1829	72.995	226.573	299.568
<b>Average</b>	33.87	1351.75	4195.79	5547.55

As can be seen in Table 1, we did not work with a huge corpus. In fact, this is a work in progress. We started with a small quantity of data, but one of our objectives is to annotate more data in order to be able to use a bigger corpus and compare the results. After having collected the corpus, we labelled it using some of the EmotiBlog elements presented in Table 2.

Table 2: Annotated elements

Element	Attribute
Polarity	Positive, negative
Level	Low, medium, high
Source	name
Target	name

As we can see in Table 2, we decided to select only a few of the elements in EmotiBlog [12]; each of them has been chosen with a special purpose. Firstly, we discriminated between objective and subjective sentences, and after that, we took into consideration only the subjective sentences with the elements presented in the table. Each of the elements indicated in the table above has been selected

<sup>2</sup> <http://www.nist.gov/tac/>

because they provide important information that is relevant to the task at hand. The polarity has the function of indicating if the opinion expressed in the sentence is positive or negative. Moreover, we labeled the data at the opinion level, choosing the level of polarity intensity between low, medium or high. Finally, we specified the source of the discourse in order to be able to detect who said what, and the target of the sentence, so as to understand the topic of the discourse. We decided not to include all the elements of EmotiBlog to avoid noise. The result of the annotation process is a gold standard which will be used to evaluate some of the aspects of the generated summaries. The subjective sentences are annotated with polarity, the level of this polarity and also with the source and the target of the discourse.

**Figure 1: Example of labeling**

```

<topic>economic situation</topic>
<topic2>government</topic2>
<topic3>banks</topic3>
<new> Saturday, May 9, 2009 My aim in this blog has largely been to
give my best and most rational perspective on the reality of the
economic situation. I have tried (and I hope) mostly succeeded in
avoiding emotive and partisan viewpoints, and have tried as far as
possible to see the actions of politicians as misguided. Of late, that
perspective has been slipping, for the UK, the US and also for Europe.
<phenomenon gate:gateId="1" target="economic crisis"
degree1="medium" category="phrase" source="Cynicus
Economicus" polarity1="negative" >I think that the key turning
point was the Darling budget, in which the forecasts were so optimistic
as to be beyond any rational belief</phenomenon>...

```

Figure 1 is an example of annotation. We would like to stress upon the fact that we indicate more than one topic. We decided to contemplate cases of multiple topics only if they are relevant in the blog. In this case, the main topic is the economic situation, while the secondary ones are the government and banks.

After having defined the topics, the first paragraph contains objective information and thus, we do not label it; we therefore annotate the following sentence that contains subjective information. As you can see, the economic crisis is the target. Finally, the polarity of the sentence is negative, the intensity level of this polarity is medium and the author is Cynicus Economicus.

#### 4.1 Annotation problems

During the annotation process we faced some difficulties, to which we tried proposing possible solutions.

The first obstacle we detected consisted in finding the topic of each blog. We started with the assumption that generally the title gives the idea of a topic, but, after having read the posts, we realized that the topic is not just the one included in the idea of the title. Furthermore, it is very usual that the

author of the new writes about a topic, but during the discussion in the blog, people change the topic of conversation. In order to overcome these problems, we decided to insert more than one topic, given that they are relevant to the global discourse. There are also blogs where no specific topic is addressed and where people talk about many different subjects and express opinions on each of them.

## 5. Generating summaries from posts

In order to produce summaries from blogs, and, more specifically, from the posts about news, we used, as a core for the summarization process, the summarization approach proposed in [author's reference]. However, as this system produces generic summaries, the blog posts had to be pre-processed and classified according to their polarity before producing the final summaries. Therefore, two sub-tasks can be distinguished within the whole process: sentence polarity classification and summary generation.

### 5.1 Sentence polarity classification

The first step we took in our approach was to determine the opinionated sentences, assign each of them a polarity (among positive and negative) and a numerical value corresponding to the polarity strength (the higher the negative score, the more negative the sentence and similarly, the higher the positive score, the more positive the sentence). Given that we are faced with the task of classifying opinion in a general context, we employed a simple, yet efficient approach, presented in [25]. At the present moment, there are different lexicons for affect detection and opinion mining. In order to have a more extensive database of affect-related terms, in the following experiments we used WordNet Affect [22], SentiWordNet [23], MicroWNOp [24]. Each of the employed resources were mapped to four categories, which were given different scores: positive (1), negative (-1), high positive (4) and high negative (-4). As shown in [25], these values performed better than the usual assignment of only positive (1) and negative (-1) values. First, the score of each of the blog posts was computed as sum of the values of the words identified; a positive score leads to the classification of the post as positive, whereas a final negative score leads to the system classifying the post as negative. Subsequently, we performed sentence splitting using Lingpipe<sup>3</sup> and classified the obtained sentences according to their polarity, by adding the individual scores of the affective words identified. As it has been shown in [25], some resources tend to over classify positive or negative examples. Thus, we have used the combined resources, which have proven to classify in a more balanced manner [25]. The measure of the intensity of the scores can also be used as an indication

<sup>3</sup> <http://alias-i.com/lingpipe/>

of the sentence importance and can thus constitute a criterion for summarization, as shown in [16].

## 5.2 Summary generation

Once all subjective sentences have been classified, we grouped them according to their polarity, distinguishing between positives and negatives. It is worth mentioning that, although the polarity of all blog sentences was determined, we only took into consideration the ones belonging to the comment posts and not in the initial news post of the blogs. This was motivated by the fact that the purpose of our summaries is to contain opinions stated by the users who have already read that news and want to express their thoughts in relation to it.

One of the main problems of blogs as far as a type of document is concerned, is the big amount of noisy information they contain. This fact can affect the quality of final summaries, and in order to avoid this, we decided to run a pre-process step, removing all unnecessary information. The problem is how to determine which information is necessary and which is not. For the purpose of our experiments, we decided that the person who stated the opinion as well as the date and time the post was written would be considered as noisy information. In some particular cases, it would be interesting to keep this information so that different strategies for grouping opinions and presenting the summary could be taken into account, such as the analysis of all the opinions of the same person. At the moment, we are more interested in subjective sentences, so that we can summarize them to provide users with the main opinions about a topic. Another problem found was the difficulty in detecting noisy information from the blogs, since each one of them presents the information in different formats. For example, regarding the authors of the posts we can find fragments such as "Paul said...", "drpower said 2:05PM on 5-13-2009", "# Julie May 14, 2009", or "Adrian Eden - May 14th, 2009 at 8:43 pm PDT". To tackle this problem, we decided to analyze the set of blogs we had and detect how the unnecessary information we wanted to remove was written; as a consequence, several manual rule-based patterns could be designed to identify this information. Having all sentences without noisy information, the next step was to run the summarization approach. It is worth mentioning that the blogs may contain orthographic and grammatical errors, which may also affect the quality of the final summaries. However, we decided not to correct them in order to maintain all the features of this kind of emerging genre. This approach employs textual entailment to remove redundant information, and computes word-frequency and noun-phrases length to detect relevant sentences within a document. The output of the system is an extract, which means that the most important sentences are extracted to produce the final summary. More specific details about the features of the summarization approaches

can be found in [21]. Two different summaries were produced for each blog, one with the positive opinions and one with the negative ones. Finally, as a post-processing stage, we bound together the summaries belonging to the same blog to produce the final summary. In the end, we generated 51 opinion summaries from different topics (economy, science and technology, cooking, society and sport), one corresponding to each blog of the corpus described in the previous sections.

## 6. Evaluation

The evaluation of summaries is a difficult task. On the one hand, automatic systems for evaluating summaries require reference summaries written by humans, and this is a very time-consuming task. Moreover, different humans would produce diverse summaries, resulting in several possible correct summaries as gold standard, making this fact another problem for the evaluation. In [26] it was shown how the result for a summary changed depending on which human summary was taken as reference for comparison with the automatic one. This problem was also presented in [27] and [28]. More recently, in [29] they stated the need of performing a more qualitative evaluation rather than a quantitative one, since summaries must contain relevant information, but at the same time, they should have an acceptable quality in order to be useful for other tasks or applications. In the DUC<sup>4</sup> and TAC conferences, summaries are evaluated manually taking into account several linguistic quality criteria, such as grammaticality or structure and coherence, for example. In this paper, we have adopted a similar approach for evaluating the generated summaries. We focus more on the quality of the summaries rather than on its content, since the content would depend on the specific need a user has at a particular moment; this has not been taken into consideration yet in our approach. However, for future work, it would be interesting to study and analyze how to produce different summaries depending on a user's profile. The criteria proposed for evaluating the opinion summaries are the following: redundancy, grammaticality, focus and difficulty. Redundancy measures the presence of repeated information in a summary. Grammaticality accounts for the number of spelling or grammatical errors that a summary presents. Focus evaluates whether it is possible or not to understand the topic of the summary, that is, the main subject of the text; and finally, difficulty refers to the extent to which a human can understand a summary as a whole or not. As can be seen, we took as a basis the criteria proposed in DUC and TAC conferences, except from the difficulty criteria which is non-conventional. We decided to contemplate this criterion, because it could be a method to evaluate the overall summary. For each one of them,

---

<sup>4</sup> <http://www-nlpir.nist.gov/projects/duc>

three different degrees of goodness were established. These were non-acceptable, understandable and acceptable. In this classification, acceptable means that the summary meets the specific criterion and therefore is good, whereas non-acceptable would mean that the summary would not be good enough with respect to a criterion. When measuring difficulty, the summaries were classified with regard to high, medium and low, being low, the better. When we evaluate the summaries with this criterion, some factors must be taken into account. The first one is the grammatical correctness; the length of the summary is another relevant element, because in fact, it is more difficult to evaluate big summaries than short ones, although longer summaries become more clear in content and understandable than short ones, as demonstrated by the results obtained. The third one is the topic. We consider as good summaries only those where the topic is clear through the text and finally, the last element is the background of the supervisor. We are convinced that evaluating a summary manually could be a very subjective task because it depends on the different backgrounds the evaluators have. The higher their level is, the clearer the summary will be.

The evaluation has been manually carried out by two potential users who, although not experts in evaluating summaries, would be very interested in having such an application to process what people think about a specific topic.

While revising the summaries, we noticed some recurrent mistakes. The first one is the punctuation; in some cases we noticed some commas missing or instead of having a comma, contain a full stop. (e.g. 'So. One option...') Also, in some cases, apostrophes are missing, in examples such as 'don't'.

The second is that sometimes we find 'PDTAh, yea'h, for example; this is the result of regular expressions that have not been processed correctly.

The third error is that in some cases the summaries start with a sentence containing a coreference element that we cannot resolve, because the antecedent has been deleted or sentences that imply some concept previously mentioned in the original text that have not been selected.

It is also worth mentioning that some of the grammatical errors are due to users' misspellings, for example 'I thikn'.

Finally, we also found some void sentences, that do not contribute to the general meaning of the summary as for example, 'I m an idiot', 'Just an occasional visitor', or 'welcome back!!!'. The tables below shows the results obtained:

**Table 3: results of the evaluation for 10% compression ratio**

	Non Accept.	Understand	Accept
<b>Redun.</b>	26%	45%	29%

<b>Gramm.</b>	4%	22%	74%
<b>Focus</b>	33%	43%	24%

**Table 4: results of the evaluation for 15% compression ratio**

	Non Accept.	Understand	Accept
<b>Redun.</b>	0%	6%	94%
<b>Gramm.</b>	2%	27%	71%
<b>Focus</b>	26%	29%	45%

**Table 5: results of the evaluation for 20% compression ratio**

	Non Accept.	Understand	Accept
<b>Redun.</b>	4%	10%	86%
<b>Gramm.</b>	0%	55%	45%
<b>Focus</b>	14%	47%	39%

**Table 6: results for the difficulty parameter**

	High	Medium	Low
<b>10%</b>	35%	28%	37%
<b>15%</b>	18%	35%	47%
<b>20%</b>	8%	51%	41%

As you can see in these tables, we decided to create summaries at three different compression ratios (10%, 15% and 20%), in order to analyze the impact of the size of a summary. The compression ratio can be defined as how much shorter the summary is with respect to the original document and it can be computed dividing the length of the summary by the length of the source text [30]. The different summary sizes would allow us to draw conclusions about the length of the summary and the qualitative evaluation. Figure 2 shows an example of generated summary for the blog 29 with a compression ratio of 10 %.

**Figure 2: an example of 10% ratio summary**

<p>Clothilde, I love the wallpapers!</p> <p>They keep everything tasty and fresh!</p> <p>Thanks a lot for the gorgeous calender desktop background.</p> <p>What a great idea and beautiful photo.</p> <p>I've just started recreating some of the easier and more attainable recipes.</p> <p>Another lovely calendar! Clotilde, have you discontinued your "Bonjour mois" newsletter?</p> <p>I'm terribly late this month but was enjoying the cheese so much that I just forgot! The peas are another winner of course.</p> <p>My only quibble would be about the name.</p>
--



The figure above is an example of automatic summary. As it can be seen, only opinions have been considered and these are presented grouped into positives, on the one hand and negatives, on the other. We considered it as good due to the fact that there are no objectives or useless sentences. The system presents subjective sentences with an emotional charge, and as a consequence this summary meets our purposes.

As you can see, the first part of the summary is composed by positive opinions and the last part by negative ones. The negative part starts with the sentence “My only quibble would be about the name”. You could notice some spelling mistakes, which are contained in the initial blog posts. Therefore, we consider as necessary to include in our system a spelling corrector in order to avoid such mistakes.

## 6.1 Discussion

Analysing the results obtained, we can draw a set of interesting conclusions. As far as the grammaticality criterion is concerned, the results show a decrease of grammaticality errors as the size of the summary lowers. We can see that the number of acceptable summaries varies from 74% to 45%, for a compression ratio of 20% and 10%, respectively. This is obvious, because the longer the summary, the more chances are for it to have orthographic or grammatical errors. Due to the informal language used in blogs, we thought *a priori* that summaries would contain many spelling mistakes. Contrary to this thought, generated summaries are quite well-written, only 4% of them, at most, being non-acceptable. Another important fact that can be inferred from the results is related to how the summaries deal with the topic. According to the percentages shown in the tables presented previously, the number of summaries that have correctly identified the topic and have therefore been evaluated as acceptable, changes considerably with respect to the different summary sizes, increasing when we change from 10% to 15%, but decreasing when changing from 15% to 20%. However, as a general trend, we can see that when taking into account the number of summaries that have not performed correctly in the focus parameter, there is a decreasing trend, reducing the incorrect summaries from 33% to 14%. This means that for longer summaries, the topic may be stated along the summary, although not necessarily in the beginning of it, whereas for shorter summaries, there is no such flexibility, and as a consequence, if the topic does not appear in the beginning, the most probable thing is that it does not appear in the summary at all. Finally, regarding redundancy, results are not conclusive, since they experiment variations in size and degree of goodness, so we cannot establish any trend. What can be seen from the results is that the summaries of 20% size obtain the best results on average over the rest of the size experimented with. This is due to the fact that this compression ratio achieves higher percentage (for the understand and accept

degrees of goodness) in two (grammaticality and focus) out of the three criteria proposed. Only the 15 % compression ratio summaries obtained better results in the redundancy criterion.

On the other hand, as far as the difficulty criteria concerned, results are also encouraging. According to the evaluation performed, the longer the summaries, the easier they are to understand in general. Grouping the percentages of summaries, we obtained that 65%, 82% and 92% of the summaries of size 10%, 15% and 20%, respectively, have, either medium or low level of difficulty, which give us an idea of they could be understand as a whole without serious difficulties. Again, for this criterion, the 20% summaries achieve the best results; this has also been proven by previous researches, which demonstrated that this compression ratio is more suitable for an acceptable quality of summaries [31]. It is worth mentioning that this criterion is rather subjective and depends to a large extent on different factors, such as the knowledge the person who reads the summaries, the number of grammatical errors the text contain, or the connectedness of the sentences. Moreover, it is reasonable to think that long summaries can be more difficult to understand, but our experiments show that is it actually the other way around, because longer summaries may contain more information than short ones, which allows the user to have more awareness of the content and what the summary is about.

## 7. Conclusion

In this paper we collected a corpus of blogs together with the comments given on them. This is an English corpus about five topics: economy, science and technology, cooking, society, and sport.

After having collected the corpus, we labeled it using a partial version of *EmotiBlog* [12], an annotation scheme for non-traditional textual genres. Furthermore, we automatically classified the polarity at sentence and also at a document level. Finally, we proposed a method for automatic summarization of similar opinions grouped for topics. The result is a summary of positive and negative opinions, divided according to their corresponding polarity. We decided to generate three different ratio summaries: 10%, 15% and 25%. In fact depending on the user’s profile a different size of summary could be more convenient than another one.

We evaluated summaries taking into consideration different parameters: redundancy, grammaticality, focus and difficulty, obtaining encouraging results.

There is no doubt about the fact that opinion summarization is a challenging task. For this reason, as future work we would like to improve our method in order to obtain better summaries. The first step would consist in evaluating our work using summaries made by humans; this is a very time consuming task, however it is

fundamental in order to assure the quality of our results. Furthermore, we would like to integrate some coreference resolution systems that could improve the quality of the language of summaries; we have some cases of noun repetitions, or in other cases, there is a sentence with a pronoun and we do not have the antecedent in the text. Another interesting challenge would be the automatic topic detection throughout the thread. Finally, we would also like to employ our techniques to other languages, such as Spanish and Italian.

## 8. Acknowledgements

The research of Elena Lloret has been supported by the FPI grant (BES-2007-16268) from the Spanish Ministry of Science and Innovation, under the project TEXT-MESS (TIN2006-15265-C06-01). The research has been also partially founded by the European Commission under FP6project QALL-ME number IST-033860.

## 9. References

- [1] P. Bo and L. Lee. Opinion Mining and sentiment analysis. Foundations and trends R. In Information Retrieval Vol. 2, Nos. 1-2 (2008) 1- 135, 2008.ingustics, and Speech Recognition. Prentice Hall, New Jersey, 2000.
- [2] K.R. Scherer. Toward a Dynamic Theory of Emotion: The Component Process Model of Affective States. Geneva Studies in Emotion and Communication 1: 1-98. 1987.
- [3] K.R. Scherer. Appraisal Considered as a Process of Multi-Level Sequential Checking, in K.R. Scherer, A. Schorr and T. Johnstone (eds) Appraisal Processes in Emotion: Theory, Methods, Research, pp. 92-120. New York and Oxford: Oxford University Press. 2001.
- [4] K.R. Scherer. What are emotions? And how can they be measured?" Social Science Information. 44(4), 693-727. 2005.
- [5] Robert E. Thayer. Calm Energy: How People Regulate Mood With Food and Exercise. Oxford University Press (New York, NY). 2001. ISBN 0-19-513189-4.
- [6] M. Tavosanis. Linguistic features of Italian blogs: literary language. New Text. Wikis and blogs and other dynamic text sources, pp 11-15, Trento, Vol. 1, 2006.
- [7] C. S. Corvalán. Sociolingüística y gramática del español. Washington DC: Georgetown University press, 2001.
- [8] H. Qu, A. La Pietra and S. Poon. Classifying Blogs Using NLP: Challenges and Pitfalls. AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs 2006.
- [9] C. Yang, K. Lin, H.-H. Chen. Emotion Classification Using Web Blog Corpora. Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence Pages 275-278 Year of Publication: 2007 ISBN: 0-7695-3026-5.
- [10] L.E. Holzman, W.M. Pottenger. Classification of Emotions in Internet Chat: An Application of Machine Learning Using Speech Phonemes. 2003. Lehigh CSE 2003 Technical Reports.
- [11] N. Kobayashi, K. Inui, Y. Matsumoto. Extracting Aspect-Evaluation and Aspect-Of Relations in Opinion Mining. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pp. 1065-1074. 2007.
- [12] E. Boldrini, A. Balahur, P. Martínez-Barco, A. Montoyo. EmotiBlog: an Annotation Scheme for Emotion Detection and Analysis in Non-traditional Textual Genres. In Proceedings of the 5<sup>th</sup> International Conference on Data Mining. Las Vegas, Nevada, USA. 2009.
- [13] A. Balahur, E. Boldrini, A. Montoyo, P. Martínez-Barco. Fact versus Opinion Questions Classification and Answering: Challenges and Keys. In ICAI'09 - The 2009 International Conference on Artificial Intelligence. Las Vegas, Nevada, USA. 2009.
- [14] J. Conroy and S. Chlesinger. 2008. Classy at TAC 2008 metrics. In Proceedings of the Text Analysis Conference (TAC).
- [15] T. He, J. Chen Z. Gui and F. Li. Ccnu at TAC 2008: Proceeding on using semantic method for automated summarization. In Proceedings of the Text Analysis Conference (TAC). 2008.
- [16] A. Balahur, E. Lloret, O. Ferrández, A. Montoyo, M. Palomar and R. Muñoz. The dsiuaes team's participation in the TAC 2008 tracks. In Proceedings of the Text Analysis Conference (TAC). 2008.
- [17] A. Bossard, M. Génereux and T. Poibeau. Description of the lipn systems at TAC 2008: Summarizing information and opinions. In Proceedings of the Text Analysis Conference (TAC). 2008.
- [18] P. Beineke, T. Hastie C. manning and S. Vaithyanathan. An exploration of sentiment summarization. In Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications, J. G. Shanahan, J. Wiebe, and Y. Qu, Eds. Stanford, US. 2004.
- [19] L. Zhang, F. Jing, X-Y. Zhu. Movie review mining and summarization. In CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management. 43-50. 2006.
- [20] I. Titov and R. Mc Donald. A joint model of text and aspect ratings for sentiment summarization. In Proceedings of ACL-08: HLT, Columbus, Ohio, 308-316. 2008.
- [21] E. Lloret, M. Palomar: 2009. A Gradual Combination of features for Building Automatic Summarisation Systems. Lecture Notes in Computer Science. 12th International Conference on Text, Speech and Dialogue.
- [22] C. Strapparava, A. Valitutti. WordNet-Affect: an affective extension of WordNet. In Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, May 2004, pp. 1083-1086. 2004.
- [23] A. Esuli and F. Sebastiani. SentiWordNet: A Publicly Available Resource for Opinion Mining. In Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2006, Italy. 2006.
- [24] S. Cerini, V. Compagnoni, A. Demontis, M. Formentelli and G. Gandini. Language resources and linguistic theory: Typology, second language acquisition, English linguistics, chapter Micro-WNOp: A gold standard for the evaluation of automatically compiled lexical resources for opinion mining. Franco Angeli Editore, Milano, IT. 2007.
- [25] A. Balahur, R. Steinberger, E. van der Goot and B. Pouliquen. Opinion Mining from Newspaper Quotations. In Proceedings of the Workshop on Intelligent Analysis and Processing of Web News Content, 2009 IEEE/WIC/ACM International Conference on Web Intelligence held in conjunction with IAT'09, September 2009, Milan, Italy.
- [26] R. Donaway, K. W. Drumme and L. A. Mather. A comparison of rankings produced by summarization evaluation measures. In Proceedings of NAACL-ANLP 2000 Workshop on Automatic Summarization. 2008.
- [27] I. Mani. Summarization evaluation: An overview. In Proceedings of the North American chapter of the Association for Computational

Linguistics (NAACL). Workshop on Automatic Summarization. 2001.

- [28] A. Nenkova. Summarization evaluation for text and speech: issues and approaches. In INTERSPEECH-2006, paper 2079-Wed1WeS.1. 2006.
- [29] J. M. Conroy and H. T. Dang. Mind the gap: Dangers of divorcing evaluations of summary content from linguistic quality. In Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008). Coling 2008 Organizing Committee, Manchester, UK, 2008.
- [30] E.H. Hovy and C.-Y. Lin. Automated Text Summarization in SUMMARIST. In I. Mani and M. Maybury (eds), Advances in Automatic Text Summarization. MIT Press, 81-94. 1999.
- [31] A.H. Morris, G. M. Kasper and D. A. Adams. The effect and limitation of automated text condensing on reading comprehension performance. Information Systems Research, Vol. 3 (1) 17-35, 1992.

# Catching the news: two key cases from today

Ruslana Margova,  
Geomedia magazine  
[naruslana@yahoo.com](mailto:naruslana@yahoo.com)

Irina Temnikova,  
University of Wolverhampton,  
[i.temnikova2@wlv.ac.uk](mailto:i.temnikova2@wlv.ac.uk)

## Abstract

This paper examines a new phenomenon in the emergence of news in Internet. Two key cases have been analyzed. The first one demonstrates the emergence of news in comments under the main news; and the second demonstrates the emergence of news in information-sharing websites and their interpretation in the newspapers. The research is based on two small corpora of texts related to these two key cases. The present study proposes some guidelines for understanding the information in the dynamic context of Internet and analyzes some possible ways to extract information from these new types of texts.

**Key words:** news, comments, blogs, information extraction, intertextuality, paratext, metatext

## 1. Introduction

The common understanding is that the newest information is in the established news media. But nowadays with the development of electronic media the real news could appear in the “non-newspapers texts” like in comments under the news, in blogs or in social networks. After that the news is immediately quoted and expanded by news agencies and online newspapers.

Internet communication removes the distinction between readers and editors. Editors become readers and the readers themselves can generate news.

The eighties were marked by the philosophy of intertextuality. This pre-Internet theory gives an explanation of what has happened in Internet today: the representation of any kind of news is often made by quotations and interpretations of quotations. The problem with the originality of the text still exists and simply said the big question is where the information resides and how it is possible to catch it.

This paper is structured as follows: Section 1 introduces the new types of texts and the quoted news, Section 2 provides the highlights of the philosophical theory which is helpful for understanding the new text types, Section 3 presents the corpora used and gives the context of the real stories with some definitions of the news, Section 4 discusses ways to extract information from corpora and shows some results, Section 5

provides the conclusions and some guidelines for future work.

## 2. A piece of philosophy: Intertextuality and information

To understand and classify all new emerging text types on the Internet, this study uses as a basis the classification made by the French theoretician Gerard Genette in 1982 [2]. Genette’s construct is based on literature analysis, without considering the new phenomenon of Internet but Internet makes the observations of Genette much clearer. Genette defined five levels of “transtextuality” (which is conceived as a complexity of the phenomena related to texts). These five levels are: hypotext (the text basis), hypertext (text’s understanding), paratext (title, subtitle, intertitle, prefaces, postscripts, notes, footnotes, final notes), metatext (commentaries, literary critique) and intertexts (relationship between two or more texts, where the most explicit form is the quotation; it also includes plagiarism and allusion).

As an example, all these levels can be seen in Internet websites – the hypotext is the information which is essential for the user – the text itself, the hypertext is the user’s previous knowledge about the current topic, combined with the new knowledge acquired from the new text, the paratext is represented in all additional features like the Internet address, the images and surrounding items including banners and advertisements, the metatext is represented by all the comments and links to other sites and texts, and the intertext is all quoted or misquoted pieces of text. This study considers the news as the hypotext, the background of the news as the hypertext, the comments as the metatext, the temporal information as the paratext and the links to other news as the metatext.

The concept of intertextuality was first expressed by the Russian philosopher Mikhail Bakhtin [12] and came to prominence in the eighties thanks to Julia Kristeva, who used the term intertextuality for describing the fact that any text is constructed as a mosaic of quotations and any text is the absorption and transformation of other texts [7]. In the present paper

we consider that “The concept of intertextuality is based on the notion that texts cannot be viewed or studied in isolation since texts are not produced or consumed in isolation. All texts exist and must be understood in relation with other texts” [1]. The two analyzed cases prove this phenomenon: the first one shows how the comments start from the original article and produce another piece of news; the second one shows how the news published by a news agency is a quotation of information taken from other websites. In fact, unofficial sources of news can also add new information to stories and in this way develop officially known stories or even introduce a new story unknown to the official news agencies. Using unofficial sources could also help to avoid copyright issues which are a burden for all the researchers collecting corpora.

Nowadays, the new types of texts are starting to be considered as a source for news. A well-known Information Security expert, Nitesh Dhanjani demonstrated at the last Black Hat computer-security conference a tool that can search for particular keywords (such as "fire" and "smoke") in posts on Twitter in order to provide an early warning for emergency responders [8].

The present paper shows that and how the new types of online texts, such as blogs, information-sharing websites or comments under the official news can be used in NLP tasks such as information extraction, anaphora resolution and text summarization.

### **3. The real stories and the corpora**

#### **3.1. First case: the story**

The owner of a Bulgarian newspaper publishes on the website of his own newspaper “the news” that he wants an official apology from a television channel because of an insult. Later, in the comments under this main news, he adds that he will receive the apology. The important point is that a person who owns a newspaper and is able to use it prefers to use the comments under an online news article like a blogger, and not his own newspaper, in order to publish a particular piece of news.

#### **3.2. Second case: the story**

After the presidential elections in Iran in mid-June, a girl (Neda) is killed during the protests in the street and her death is filmed by bystanders and broadcasted over the Internet – in Youtube ([www.youtube.com](http://www.youtube.com)) and social networks like FaceBook ([www.facebook.com](http://www.facebook.com)) and Twitter ([www.twitter.com](http://www.twitter.com)). The news is later reported by all daily online newspapers and the social

networks and the places where it first appeared are quoted as the official sources.

#### **3.3. Two corpora**

Two small corpora have been developed for these two different cases. The first one (10340 words) represents news published by editors and the postings of bloggers under it. The posting labels – like the time markers, the bloggers’ names and the bloggers' moods - are preserved in the corpora for ease of future processing. The corpus consists of the official news and 186 comments.

The second corpus (21659 words) is created from three different types of texts – the videos posted on an information-sharing website ([www.youtube.com](http://www.youtube.com)), a blog (of the writer Paulo Coelho), and the results of the Google News Search engine (which retrieves the news published in online newspapers).

#### **3.4. The news or “there is nothing older than yesterday’s newspaper”**

In the current study the definition of 'news' is very important. In the theory of journalism 'news' is previously unknown information about a recent and important event. The problem is how to define what is previously known and what is not.

Information Extraction (IE) is the sub-area of NLP which deals with the extraction of news. In Information Extraction the first of three main tasks is to determine what are the important types of facts for a particular domain [4]. Or in other words - to define what is previously known and what needs to be known.

The other two main tasks for each type of fact are: determining the various ways in which it is expressed linguistically; identifying instances of these expressions in text.

For the task of catching the news, the definition of 'previous knowledge' is also very important. The temporal information about the comments could be used as a marker of the news. In our two corpora we keep the date and the hour from the paratext information. Furthermore we consider that the newest information, identified by the latest temporal label, could be conceived as news.

Information Extraction is the automatic identification of selected entities, relations or events in free text [3]. Event Extraction (EE) is a particular type of IE. EE can be defined as extracting all occurrences of a relationship between specific participants in an event from text. In order to capture a particular relationship between particular participants in a particular situation, patterns are being built. Usually attention is focused on identifying and extracting Named Entities (NEs), such as names of persons,



organisations and locations of time markers, people positions. EE from news articles is usually done by grouping similar articles into topic clusters based on statistical word co-occurrences.

A further step is tracking the development of news in time [10,11]. An example of news topic tracking and news linking over time is the Europe Media Monitor system (EMM) which groups around 50,000 articles per day into clusters per topic and per language and then links daily clusters over time into stories, in this way tracking news story development over time [10].

In our case IE could be considered as the technology of extracting the information and the time label could be conceived as a guarantee for the novelty of information. In the first case the previous knowledge is contained in the article published on the website. In the second case the previous knowledge is in the news articles on the web about the elections in Iran. Both cases show how previous knowledge is enriched by additional information from comments or websites and how news is created.

## 4. Text analysis and searching for the news

### 4.1. First case – general remarks.

In the first corpus there are texts with different characteristics: first - an edited journalistic text; second – the “freestyle” comments of nicknamed people. An important point is that the post-editing process is easier in online media – every news article can be improved in real time. So in our case we analyze the first state of the article before the emergence of comments and its further rewrite.

#### 4.1.1. The article

Typically, the article is clearly structured, with clear simple phrases, with exact quotations and with enough information for the readers. The published article could be considered as a well written journalistic text meeting all requirements. The news is in the first phrase and additional information follows.

#### 4.1.2. The comments

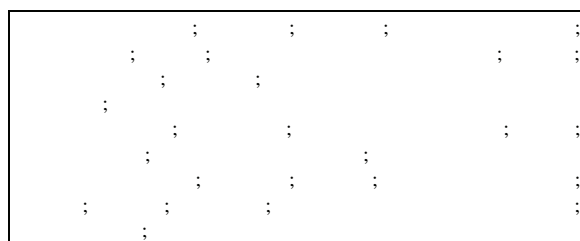
Grammatically the comments could be distinguished by the faults – bad constructions, unfinished phrases, etc. Also, the comments are posted by nicknamed people. The content of the comments can also be distinguished by the fact that it is emotional, defending two main opposite positions – of the newspaper owner or of the guest of the television channel.

Information extraction and Opinion mining could help to find more information about the main

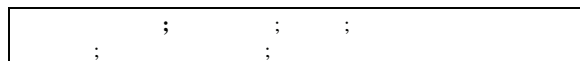
participants in the story. For example the extraction of named entities and the relationships between entities could provide some additional information – who is who, who is liked by whom, what somebody did or didn't do, etc. But all that additional information will not be the news, because we consider it to be previous knowledge.

The names of the newspaper owner and the guest of the television channel appear more often and are simple to identify. There are also other people, firms and organisation names mentioned. An interesting question is the name co-reference resolution which could help for future opinion mining.

Below follows an example of the list of proper names and co-references which are mentioned in the main news text and used by the bloggers:



Another example is the list of proper names and co-references which do not appear in the news texts but are used by the bloggers:



A further in-depth linguistic analysis may help the development of rule-based Information extraction of comments. Information extraction would improve the general knowledge about these people and would facilitate the better understanding of the problem.

#### 4.1.3. Newsmakers or bloggers, or both.

In one of the postings, a blogger puts his initial as the TV guest, and in this way identifies himself as the representative of the television programme guest. His comment is a quotation of the letter of the TV guest, sent as an answer to the accusations of the newspaper owner.

The linguistic analysis of the text shows that the first phrase of the posting is not well written from the point of view of the typical journalistic style – there is an inversion in the noun phrase, missing information (to whom/where), ambiguity. The second phrase starts with a repetition – which is also a fault of style in Bulgarian. The rules of quotation are also not applied.





[...],

The quoted text is well edited and structured. It is almost an exact quotation of the interview taken from the television broadcast. The deep content analysis of that text and the transcript could prove its authenticity. Unfortunately the intervention of the guest in the comments is still not news – it is only a comment. The bad style in the first phrase makes unclear where/to whom the guest sent this letter, and how this blogger knows about that – so whether it is true or not.

A little bit later, as the temporal labels show, another blogger identifies himself as the owner of the newspaper and adds a piece of news in the comments – he announces that he will participate in two television shows to explain the situation and that he already has the promise of the director of the television channel that there will be an apology.

"

The news is introduced by the phrase: “ (the breaking news of the last hours is). Here we have some new information, said in first person, from somebody who claimed to be the real owner.

After this comment, that announcement is repeated and quoted by many media. (And finally it really happened.)

The linguistic analysis of this text shows that it’s not an edited text – there are two repetitions ( ) and one misspelled word ( ). But the important thing in this comment is the news – the new development of the event – that there will be an apology.

As was already mentioned, the news in the website can be easily post-edited. The news about this apology is post-edited in the same website but after the comment of the owner, and not before. This shows that the news is really born in the comments.

#### 4.1.4. How much news in the corpus

We mentioned the importance of the introductory phrase. In the corpus of comments there are only two other postings, conceived as news from the bloggers. All the other texts are interpretations, analysis and opinions.

The first comment, which is conceived as news from a blogger, is formed by an introductory phrase:

:

and the quotation of the news:

” 26,1% “ ” ” ”

This news is real and also published in the media, but here it is like a comment. It will be news for all the bloggers who don’t know it, but it’s not news emerging in these comments. The other news is introduced by the paratext – the nickname of the person is (the news).

1.

In this case the blogger reproduces in the comment the announcement of the owner, made a little bit earlier. This case could be an example of the third main task of IE – the identification of instances of important expressions.

The analysis of the corpus shows that the **real news is not very often seen in the comments**. In all 186 comments there is only one, **which is introduced by a concrete clear phrase with temporal identification**. The other two examples show different ways of introducing new information. Further work with other corpora could make clear how news is introduced.

In the case shown, the intervention of the owner and his announcement was easy to identify, thanks to the paratext markers (time, nickname).

But there are still many different problems here: whether everyone who presents himself as a particular person is really the same person; how the reader could be sure about that; and furthermore - how the automatic extraction of information has to be made and how authentic it could be. Another question is how often such a kind of event development is possible.

## 4.2. Second case – general remarks

### 4.2.1 Sources and confusions – youtube.com, google news, Associated Press, blogs

As has been already mentioned, for the second analyzed case, the previous knowledge about the situation in Iran is taken from articles and all kinds of

web resources. In this part of the study we make an attempt to reconstruct the emergence of news from the web, but not from an official news agency source and news websites. The temporal anchor is June 21 – the day when the video of Neda’s death appeared in shared information websites. At the present moment there is no exact answer about where the video was first published – on Youtube, in Facebook or in Twitter, but the important fact is that the video appeared first in a non-news website and the question is again **how to catch the news from the Internet**.

#### 4.2.2. Event reconstruction and new channels

The event reconstruction will give some clues. The search in news.google.com by keyword “iran” on 21 June doesn’t give the result of Neda. Today these two words are strongly co-related.

The video of Neda’s death was multiplied in the postings in the video-sharing websites and, thanks to an unknown editor, emerged as news. The huge problem is how to catch such news and the possible answer is to change our perception and to start considering these social networks and websites for shared information as **another type of news channel**.

The news published by Associated Press on 22 June 2009:

CAIRO (AP) — Amateur video of a young Iranian woman lying in the street — blood streaming from her nose and mouth — has quickly become an iconic image of the country’s opposition movement and unleashed a flood of outrage at the regime’s crackdown.

The footage, less than a minute long, appears to capture the woman’s death moments after she was shot at a protest — a powerful example of citizens’ ability to document events inside Iran despite government restrictions on foreign media and Internet and phone lines.

The limits imposed amid the unrest over the disputed June 12 election make details of the woman’s life and events immediately preceding her apparent death difficult to confirm. But clips of the woman being called Neda are among the most viewed items on YouTube — with untold numbers of people passing along the amateur videos through social networks and watching them on television.

The images entered wide circulation Saturday when two distinct videos purporting to show her death appeared separately on YouTube and Facebook.[...]

Thousands of people inside and outside Iran have written online tributes to the woman, many condemning the government and praising her as a martyr. Some posted photos of a gently smiling woman they said was Neda, some calling her “Iran’s Joan of Arc.”

The first representation of the news of Neda’s death in the news agencies shows some interesting facts:

- Even Associated press is retelling the story of Neda’s death seen in amateur videos –even highly reputable media are obliged to use in that case an unverified information channel.

- The Associated press quoted YouTube and Facebook as sources – so these two channels are considered to be information factors.
- The Associated press quoted people inside and outside Iran who have written online tributes to the woman – web communication is conceived as witnesses’ stories.
- It is difficult to confirm the event – from the point of view of journalism’s ethics – where the truth is – this fact has to be kept in mind.

After the acceptance of these channels as news information media, the next step is how to use these new channels. We present some of the possible perspectives of using this new type of media.

#### 4.2.3. Summarization of the information in Youtube.

The search in www.youtube.com of the keyword “neda” returns a set of titled videos. The collection of all these titles could be conceived as a kind of summarization of the real story about the death of Neda Agha-Soltan.

The collection of titles:

Neda’s Death Becomes Iranian Symbol  
 Neda Agha Soltan, killed 20.06.2009, Presidential Election Protest, Tehran  
 Her name was Neda  
 Neda before she gets shot  
 IRAN PROTEST IMAGES IN TRIBUTE TO NEDA  
 RIP Neda Soltani, Neda! Don't Be Scared, Neda!  
 CNN: "Death Of Neda" Video Becomes Symbol Of Iranian Protests  
 Fiance tells of Neda's last moments - 23 Jun 09  
 Twitter Revolution – Iran  
 CIA KILLED NEDA  
 United for Neda  
 I Am Neda  
 For Neda  
 John McCain Addresses Killed Iranian 'Neda' on Senate  
 SONG FOR NEDA original music by Greg V. In honor of Neda

The manual reconstruction of the extracted titles gives almost the full story:

Neda Agha Soltan is killed on 20.06.2009, in a presidential election protest at Tehran. Her fiance tells of Neda’s last moments. She becomes an Iranian symbol. There are some allegations that Neda is killed by CIA. There is a Twitter Revolution – Iran. There is a tribute to Neda – songs, addresses in the Senate.

Such kind of work could be done for many themes. Thus, further automatic summarization over the results of the search for some key words can be helpful for the creation of full information about a concrete topic. Once different titles of the same video are recognized to refer to the same story, coreferential chains, and lists of co-referents relating to the same term can be easily built. This kind of list can be constructed also from the different ways of naming the same personages involved

in an event by the different people commenting under the news. Collecting such lists could also help anaphora resolution in news topic tracking.

#### 4.2.4. Additional information from blogs

Additional information about some topics could be found also in blogs. The videos of Neda's death are posted on many private websites and blogs and we tried to choose randomly one of them which has many comments in order to make an example of this possibility. The chosen blog is of the famous Brazilian writer Paulo Coelho. The post is from 23 June 2009 and his remarks are:

My best friend in Iran, a doctor who showed me its beautiful culture when I visited Teheran in 2000, who fought a war in the name of the Islamic Republic (against Iraq), who took care of wounded soldiers in the frontline, who always stood by real human values, is seen here trying to resuscitate Neda - hit in her heart.

The obvious conclusions from this piece can be two – first, the owner of the blog knows one of the persons in the video – the doctor who tries to help the girl to survive (in the blog the name of the doctor (Arash Hejazi) is published on June 26<sup>th</sup>, 2009 and there is also an exchange of correspondence between the doctor and the blog's owner); second, the video could be true. No one has announced the name of the doctor until that moment and no one has proved that the video is real. Five days later the news agencies announced: *Arash Hejazi is wanted by intelligence ministry and Interpol*.

In this case the IE can be used as a technology for catching important information and could be enriched with time labels – to identify the news. In both cases – the newspaper owner's story and Neda's death story - the number of comments and the activity of bloggers show that these two news are important (the first - for Bulgaria, the second – for all over the world).

## 5. Conclusions and further work

The paper aims to show some of the perspectives of news emerging in shared web texts. Starting from the understanding of intertextuality, we presented the development of the news in two key cases. The main conclusion is that the shared-information websites, social network sites, blogs and comments under a particular article can be without a doubt conceived of as new news channels.

The Information Extraction technology can help as a method of catching suitable information from these new sources. The introduction of time labels as a guarantee of the novelty of information is helpful to determine which is real news.

Future work will start with in-depth linguistic analysis of the news comments and blog posts to show their differences from classical news articles texts. Considering news article texts as a form of a controlled language will help to make this distinction clearer. The collection of a more representative corpus of new text types and their analysis should be considered as the next step.

## 6. References

- [1] Franklin, Bob, et al., 2005, Key concepts in journalism studies, Sage Publication.
- [2] Genette, G., 1982, *Palimpsestes. La littérature au second degré*, Paris: Éditions du Seuil.
- [3] Grishman, R., 2003. Information Extraction. The Oxford Handbook of Computational Linguistics. Chapter 30. Edited by R. Mitkov. Oxford University Press.
- [4] Grishman, Ralph, 2005, NLP: An Information Extraction Perspective, Recent Advances in Natural Language Processing IV, John Benjamins Publishing Company, edited by Nicolas Nikolov, Kalina Bontcheva, Galia Angelova, Ruslan Mitkov.
- [5] Howard, Rebecca Moore, 2007, Understanding "Internet plagiarism", The Writing Program, Syracuse University, USA.
- [6] Kovach, Bill, Rosenstiel, Tom, 2001, The Elements of Journalism: What Newspeople Should Know and The Public Should Expect, Three Rivers Press.
- [7] Kristeva, J. 1978, *Semiotike Recherche pour une Sémantique*, Edition Points.
- [8] Naone, E. Mining Social Networks for Clues. Technology Review July 31, 2009.
- [9] Pfeifle, Mark, 2009, A Nobel Peace Prize for Twitter?, The Christian Science Monitor.
- [10] Pouliquen Bruno & Ralf Steinberger (2008). Story tracking: linking similar news over time and across languages . In Proceedings of the 2nd workshop Multi-source Multilingual Information Extraction and Summarization (M=IES'2008) held at CoLing'2008. Manchester, UK, 23 August 2008.
- [11] Richter, M. Analysis and Visualization for Daily Newspaper Corpora. Proceedings of RANLP, (2005) 424-428.
- [12] , . . ., 1929, . . .
- [13] [http://www.dnevnik.bg/bulgaria/2009/07/22/759314\\_ivo\\_proko\\_piev\\_poiska\\_bi\\_ti\\_vi\\_da\\_mu\\_se\\_izvini\\_zaradi/](http://www.dnevnik.bg/bulgaria/2009/07/22/759314_ivo_proko_piev_poiska_bi_ti_vi_da_mu_se_izvini_zaradi/)
- [14] <http://paulocoelhoblog.com/?s=iran>
- [15] <http://www.ap.org/>
- [16] <http://news.bbc.co.uk/>

# Detecting Opinion Sentences Specific to Product Features in Customer Reviews using Typed Dependency Relations

Ashequl Qadir  
University of Wolverhampton  
Stafford Street, Wolverhampton  
West Midlands, WV1 1SB, UK  
ashequl.qadir@wlv.ac.uk

## Abstract

Customer reviews contain opinions of the customers who purchased products and expressed opinions concerning their satisfactions and criticisms. Due to vast availability of product reviews in the web, it is extremely time-consuming and at times confusing for a new customer to manually analyze the reviews prior to buying a product. Reviews generally involve the presence of product feature specific factual information along with the opinion sentences depicting the pros and cons of a bought product. The unstructured format of the text reviews from most of the web review sources necessitates the automatic identification of opinion sentences from the customer reviews, and also the identification of explicitly visible and implicitly present product features associated with the opinion sentences. In this paper, a process has been described where typed dependency relations such as open clausal complements or adjectival complements have been utilized to identify opinion sentences specific to product features. The typed dependency relations in the identified opinion sentences are then used to associate a product feature to an opinion sentence with the help of the product feature associated frequent words extracted from a previously managed customer review corpus.

## Keywords

Product features, customer reviews, opinion sentences, typed dependency relations, frequent word association.

## 1. Introduction

After purchasing a product, customers quite often write their experiences in their reviews. These reviews contain their opinions about the product they purchased. These customer reviews are different from the traditional texts because they are written spontaneously and are small texts focused on a single topic or a product having several attributes and features. This relatively new type of texts mostly conveys sentiments about the topic or the purchased product and is getting widely popular day by day providing researchers with interests to explore a wide range of scopes and possibilities about how these texts can be processed and necessary information can be retrieved.

A new customer, before purchasing a product, quite often tends to look up the previously written reviews to analyze the positive and negative aspects of the product he intends to buy. This practice is increasing rapidly making it very important to formulate ways to process and retrieve information automatically from the text reviews. The products, for which the reviews are written, are associated with several product features, usually common to a particular

product domain. The reviews can contain very general opinions such as *'I am very happy with this product'* or can also contain product feature specific opinions such as *'It is very easy and simple to use'*, associated with a usability feature. Along with the opinion sentences, factual information such as *'it has a pink metal case'* can also be found in the reviews that do not contain any opinion of the reviewer; rather gives a factual description. As a result, before making a decision on the polarity of the opinions, it is very important to identify the opinion sentences and to identify the product features associated with them. Most of the popular products usually have many reviews written for them and it takes a significant amount of time to go through the review sentences manually in order to separate the opinion sentences from the others.

There are a number of review sources in the web where reviews can be found. E-commerce sites such as amazon, opinion sites such as epinions, forums, blogs etc are very well known sources for reviews and also very popular among the customers where reviews written by them can be found. Processing these mostly unstructured text reviews automatically is considered very challenging because of the frequent use of the informal expressions and terms, grammatically incorrect sentences, misspelled words etc. that can be occasionally found in the reviews.

Words forming a sentence have certain grammatical relations with each other based on their part-of-speech definitions, positions in the sentences etc. Some of these relations are representative of the functional features of a product for which the customers express their opinions. In this paper, a process has been described that utilizes the typed dependency relations of the words in sentences to identify opinion sentences written on product features. Because some of the relations are representative of the product features, these words are then utilized to assign a probable product feature to each of the opinion sentences under consideration. To utilize the dependency relations, Standard typed dependency relation representations [18] are chosen over PARC[20] representations because Standard typed dependency relations offers[17] more fine-grained distinctions in relations such as breaking down an unsubcategory relation into several more distinctive relations like adjectival modifiers, prepositional relations, open clausal complements etc. This helps to obtain more precise dependency relations suitable for the designated purpose.

## 2.Related Work

Opinion sentence identification has been mostly approached by the researchers by means of determining the presence of specific parts-of-speech such as adjectives, adverbs etc. or a list of seed words that may potentially represent opinions. Research of Wiebe[1] and Hatzivassiloglou et al.[2] showed that adjectives can potentially contribute towards identifying subjective sentences. Turney[3] used specific orientation of part-of-speech tags to extract phrases that can represent opinion. Godbole et al.[4], Kim et al. [5] used a small seed list of lexicons, expanded later, for their sentiment identification process. Riloff et al. [6] researched on identifying extraction patterns for subjective and objective sentences using subjective clues such as single words or N-grams. Wiebe et al.[7] worked on using word collocations that can act as subjectivity clues for identifying opinion sentences. Yu et al.[8] used the similarity between the opinion sentences within a given topic to identify opinion sentences and Naïve Bayes classification scheme to distinguish between opinion and factual sentences. Wilson et al.[9] used dependency relations of words as one of their syntactic clues for determining subjectivity strength. Fei et al.[10] researched on utilizing the dependency relations of words in sentences for a target specific sentiment extraction.

Previous research works in product feature identification were mostly focused on explicit product features only. Yi et al.[11],[12] and Liu et al[13] worked on identifying explicit product features by extracting noun phrases of specific patterns. Popescu et al.[14] utilized parts and properties of a given product to identify product features. Ghani et al.[15] approached explicit product feature extraction as a classification problem. Qadir[16] used frequent word associations learned from a previously managed corpus to associate product features with sentences. Zhuang et al.[19] utilized dependency grammar graph to mine explicit feature-opinion pairs in movie review domain.

The approach described in this paper differs from the above mentioned previous researches by using Stanford typed dependency representations[17]. Specific typed dependency relations are utilized to differentiate opinion sentences from factual ones. Words forming the specific dependency relations are analyzed with frequent product feature associated words to assign a product feature to each of the opinion sentence.

## 3.Review Collection and Pre-processing

There are several product review sources available in the web. These sources can be e-commerce sites, opinion sites, forums, blogs etc. For this experiment, 100 reviews have been collected from amazon using amazon web services. Amazon web services (AWS) allows the developers to automatically collect plain text reviews. The collected reviews are from the domain ‘*Electronics*’ and the product type is ‘*hard disk*’. 50 reviews have been used to identify the frequent words that are usually associated with the product

features. This set of reviews has been used as a training corpus.

Each of the sentences in the set of reviews has been annotated manually with product feature titles. Sentences that do not convey any opinion of the reviewer have been tagged as ‘No Opinion’ and the sentences that convey only general opinions of the reviewers and not any product feature specific opinions are tagged as ‘General’. Five other distinctive product feature titles have been identified from the reviews. Table 1 gives examples of the opinion lines that can be associated with these five different product features. These examples are taken from the collection of review texts.

**Table 1. Product feature associated opinion sentences**

<b>Product Feature</b>	<b>Opinion Sentence</b>
Usability	<i>‘It was incredibly easy to set up and use.’</i>
Design	<i>‘I like its design and the fact that I only need one cable.’</i>
Performance	<i>‘Works perfectly and is completely reliable, no problem at all.’</i>
Portability	<i>‘I found this product really useful for transport as it is that small.’</i>
Speed	<i>‘The speed and capacity of the Passport drive are impressive.’</i>
General	<i>‘A satisfying product.’</i>

The rest 50 reviews are kept for evaluating the process to identify opinion sentences and associate a product feature with each opinion sentence.

## 4.Methodology

The methodology section divides the whole process into two major tasks. To identify the opinion sentences, relevant typed dependency relations are selected and utilized. And to assign a product feature to each of the opinion sentences, frequently associated words are obtained from a previously managed corpus, normalized within the product feature scope by tf.idf metric and then utilized in the association process.

### 4.1 Finding Opinion Sentences

#### 4.1.1 Typed Dependency Selection

Stanford Typed Dependencies Manual[18] gives definition to 55 binary grammatical relations between a governor and a dependent that can possibly be present in a sentence. From them, 3 of the relations have been selected as they can indicate a probable presence of product feature specific or general opinions in review sentences.

#### 4.1.1.1 acomp - Adjectival Complement

An adjectival complement (acomp)[18] of a VP is an adjectival phrase which functions as the complement (like an



object of the verb); an adjectival complement of a clause is the adjectival complement of the VP which is the predicate of that clause. The governor component of the acomp typed dependency relation is a verb indicating a functionality of the product and the dependent component is an adjective indicating an opinion of the reviewer on that functionality. Table 2. gives examples of the components for acomp typed dependency relation taken from the review sentences of domain ‘*Electronics*’. Examples are given for the most frequent types of verb forms.

**Table 2. Example of acomp relation as opinion indicator**

Dependency Relation	Component Example	Indication
acompany	worked/VBD fine/JJ	Possible Opinion
acompany	proved/VBN reliable/JJ	Possible Opinion
acompany	works/VBZ well/JJ	Possible Opinion

#### 4.1.1.2xcomp – Open Clausal Complement

An open clausal complement (xcomp)[18] of a VP or an ADJP is a clausal complement without its own subject, whose reference is determined by an external subject. In case of xcomp typed dependency relation, verb as the governor component and adjective as the dependent component and also adjective as the governor component and verb as the dependent component have been considered. Table 3. shows the examples taken from review lines in domain ‘*Electronics*’ where xcomp can possibly indicate the present of an opinion in a review sentence.

**Table 3. Example of xcomp relation as opinion indicator**

Dependency Relation	Component Example	Indication
xcomp	easy/JJ use/VB	Possible Opinion
xcomp	rendering/VBG impossible/JJ	Possible Opinion
xcomp	found/VBD difficult/JJ	Possible Opinion
xcomp	makes/VBZ ideal/JJ	Possible Opinion
xcomp	find/VBP convenient/JJ	Possible Opinion
xcomp	experienced/VBN similar/JJ	Not Opinion

#### 4.1.1.3advmod –Adverbial Modifier

An adverbial modifier(advmod)[18] of a word is a (non-clausal) RB or ADVP that serves to modify the meaning of the word. Unlike acomp and xcomp typed dependency relations, advmod relation is less likely to indicate the presence of a product feature specific opinion because of the absence of the verb, but more likely to indicate the presence of a general opinion because of the presence of the adjective or the adverb that modifies the adjective or the verb. When the governor component is an adjective and the dependent component is an adverb, advmod mostly indicates the presence of an opinion, and such combination can be found very frequently. Also, when both the governor component and the dependent component of the advmod typed dependency relation are adverbs, it does not represent any product feature functionality by itself. On the other hand, when the governor component is a verb, advmod relation quite often does not indicate the presence of an opinion, but the verb remains an indicator of a functionality of the product for which the reviewer expresses his opinion somewhere else in the sentence. It is needed to be mentioned that adjectival modifier (amod) typed dependency relation sometimes represents opinion and sometimes does not; thus could not be used as a definitive indicator to identify product feature specific opinion sentences. Table 4. shows examples taken from review lines in domain ‘*Electronics*’ where advmod relation can possibly indicate the present of an opinion in a review sentence.

**Table 4. Example of advmod as an opinion indicator**

Dependency Relation	Component Example	Indication
advmod	well/JJ amazingly/RB	Possible Opinion
advmod	easily/RB very/RB	Possible Opinion
advmod	loads/VBD fast/RB	Possible Opinion
advmod	looks/VBZ especially/RB	Not Opinion
advmod	fits/VBZ perfectly/RB	Possible Opinion
advmod	recognized/VBN straight/RB	Not Opinion
advmod	satisfied/VBN very/RB	Possible Opinion
advmod	priced/VBN reasonably/RB	Possible Opinion

#### 4.1.2Opinion Sentence Detection

When the above mentioned typed dependency relations are present in the review sentences, following algorithm has



been used to determine whether a review sentence can be considered as an opinion sentence.

**Figure 1. Algorithm to identify opinion sentences**

1. for each sentence in review text
2.     set Opinion\_Flag=False
3.     check acomp\_presence
4.     if present
5.         if governor is any form of verb
6.         if dependent is any form of adjective
7.             set Opinion\_Flag=True
10.    check xcomp\_presence
11.    if present
12.     if governor is any form of adjective
13.     if dependent is any form of verb
14.         set Opinion\_Flag=True
15.     else if governor is any form of verb
16.         if dependent is any form of adjective
17.             set Opinion\_Flag=True
18.    check xcomp\_presence
19.    if present
20.     if dependent is any form of adverb
21.     if governor in any form of verb
22.         set Opinion\_Flag=True
23.     else if governor is any form of adverb
24.         set Opinion\_Flag=True
25.     else if governor is any form of adjective
26.         set Opinion\_Flag=True

**4.2 Assigning Product Features**

Each of the opinion sentences is assigned with a product feature with the help of the frequently associated words that appear with the selected typed dependency relations mentioned above.

*4.2.1 Counting Frequent Words*

As a product feature tag is assigned to each of the review sentences in the test data set, word counts are therefore done only within the product feature scopes. But instead of taking all the words of each sentence into consideration, only the words in component elements of the typed dependency relations are counted as they can be considered to carry the most indicative information to identify a product feature. Rest of the words in each sentence is ignored to avoid undesired words that do not relate to any specific product feature. Any word which is a function word is also ignored and is not involved in the counting process so that the common words that are present in any text can be avoided. While counting, lemmatization is used to consider only canonical form of the words so that the frequency of the

words does not get distributed over different representations of same words.

If  $N$  is the total number of review lines present in the test data set and  $p_1, p_2, \dots, p_j \in P$  is the set of product features then word frequency count,  $WC_j$  for word  $w$  within  $p_j$  product feature scope can be denoted by the following equation:

$$WC_j = \sum_{i=1}^N w_{i,j}$$

where,  $w_{i,j}$  is the frequency of the word  $w$  at review line  $i$ , associated with product feature,  $p_j$ . For different values of  $j$ , word frequency of the same word  $w$  will be different because associated product feature  $p_j$  will be different.

To include synonyms of the words in the counting process, Wordnet’s synset for each word has been used. But because each of these words in synsets was not originally present in the review sentence, there is no surety that the synonym under consideration will be appropriate under the context. In addition to that, there can be more than one synsets in case of polysemous synonyms. Therefore, instead of counting each synonym for single occurrence, each of the synonyms is divided by the total number of synonyms found from all the synsets having the original word to represent a probability measure. That is, for  $k$  synsets having  $n_i$  synonyms in each, the probability of each synonym to be the appropriate synonym of the original word,  $w$  is considered by the following probability function:

$$P(w) = \frac{1}{\sum_{i=1}^k n_i}$$

This does not eradicate the noise in the word list introduced by polysemous synonyms, but minimizes the impact. This probability score is used as the frequency of the word synonyms.

*4.2.2 Normalizing with tf.idf metric*

To normalize the word frequencies, tf.idf metric has been used. If  $WC_{i,j}$  is the frequency of word  $w_i$  in a product feature scope  $p_j$ ,  $k$  is the number of total words in  $p_j$ , then term frequency,  $tf_{i,j}$  can be denoted by,

$$tf_{i,j} = \frac{WC_{i,j}}{\sum_k WC_{k,j}}$$

if  $|P|$  is the total number of product features assigned in the corpus,  $|p : \{w_i \in p\}|$  is the number of product features with which the word  $w_i$  appears, then inverse document frequency  $idf_i$  can be calculated by the following,

$$idf_i = \log \frac{|P|}{s + |p : \{w_i \in p\}|}$$

The inverse document frequency calculation process suffers from a possibility of division by zero error. In the evaluation review data set, if there are new words that do not appear with any of the product features in the training data set, the denominator at the right side of the inverse document frequency calculation equation will have a zero value and  $idf$  cannot be calculated. To avoid this problem, a soothing parameter  $s$  has been used in the denominator. The value of  $s$  has been selected to be 0.001 which is a very small value that does not have any impact of its own in the calculation process. And finally, the  $tf.idf$  weight metric for word  $w_i$  can be calculated by multiplying term frequency  $tf_i$  with inverse document frequency  $idf_i$ .

#### 4.2.3 Assigning Product Features

For each product feature, a product feature score is calculated using the following formula:

$$PFS = \sum f(acomp) + \sum f(xcomp) + \sum f(advm)$$

where  $f(relation)$  is a function that calculates the  $tf.idf$  weight score for each of the components of a typed dependency relation considering a specific product feature.  $tf.idf$  scores for both the words at the governor and dependent position of the typed dependency relation is summed up for all the selected typed dependency relations that can be found in the sentence that is needed to be assigned a product feature. This score of each sentence is calculated for all the product feature classes.

$PFS$  represents the contribution of a set of words in a sentence towards different product features. Because the  $tf.idf$  metric yields different scores for each of the words within different product feature scopes,  $PFS$  will have different values for each of the product features. When a product feature achieves a higher  $PFS$  value than the others, this means the words in the opinion sentence under consideration are more indicative of that product feature than of others. If  $c$  is the product feature class for which the product feature score,  $PFS$  is calculated, then each opinion sentence is assigned to a product feature class  $c^*$  where,

$$c^* = \arg \max_c PFS$$

From all the  $PFS$  scores calculated, a threshold value has been selected to be 1% of the highest  $PFS$  score. This is because some of the sentences that do not contain any opinion might have few words common with sentences that contain product feature specific opinions. But if not indicative enough, these words will yield a relatively low  $PFS$  score because they do not appear very frequently with the product feature specific opinion sentences. That is why, below this threshold value,  $PFS$  score is considered to be not strong enough to indicate a product feature and thus the corresponding sentence is considered as a not opinion bearing sentence.

## 5. Results

Manually annotated review data set of 50 reviews, kept for evaluation of the system, consisted of a total of 220 sentences having 113 opinion sentences and 107 sentences with no opinion. Table 5 shows sentence and word distribution of the selected product features in evaluation review set.

**Table 5. Sentence and word distribution of test data**

Product Feature	No. of Sentences	Average words per sentence (Without Function Words)	Average words per sentence (With Function Words)
General	57	5.63	7.77
Usability	16	11.44	13.69
Design	15	6.53	8.80
Portability	9	10.22	12.89
Performance	9	9.33	11.44
Speed	7	12.57	16.00

Based on the manually annotated test set of 50 test reviews in domain 'Electronics' for product type 'hard disk', the precision and recall scores for opinion detection are presented in Table 4.

**Table 4. Evaluation score for opinion sentence detection**

Precision	Recall	F-measure
0.7231	0.4159	0.5281

The evaluation scores for the assigned product features based on the manually annotated test set of 50 test reviews in domain 'Electronics' for product type 'hard disk' are presented in Table-5.

**Table 5. Evaluation score of product feature assignment**

Product Feature	Precision	Recall	F-measure
General	0.7778	0.1228	0.2121
Usability	0.9231	0.7500	0.8276
Design	0.6364	0.4667	0.5385
Performance	0.5833	0.7778	0.6667
Portability	0.7143	0.5556	0.6250
Speed	0.3077	0.5714	0.4000
No Opinion	0.5742	0.8318	0.6794

in the evaluation scores of product feature assignment, some of the product features achieved satisfactory result. This is because different verbs represent different functionalities of a

product and assigned product feature to the opinion sentence. On the other hand, the opinion sentences that do not convey any product feature specific opinion; rather convey opinions of the reviewers in general categories are difficult to identify. As a result, the recall score is relatively low for general opinion sentences.

It has been observed that, quite often, a single sentence carries opinions about more than one product features. In this experiment, such sentences were tagged with only one product feature title. As a result, the words that are usually associated with the other product features but present in the same sentence contributed wrongly towards both. Appropriate segmentation methodology that can segment a single sentence in a way that only one product feature can be assigned to each sentence is needed to be applied in order to obtain a better result.

## 6. Conclusion

This paper discusses a process to detect opinion sentences and assigns a product feature to each opinion sentences. Typed dependency relations and frequent word associations have been utilized to achieve the desired goal. The obtained results leave room for improvement possibilities. Also, the process has been experimented within a very small scope. Future works will involve identifying appropriate segmentation methodology to aid the system, implementing the process in a number of varied domains and exploring left and right context of the dependencies for more supporting information towards product feature assignment.

## 7. References

- [1] J. Wiebe. Learning Subjective Adjectives from Corpora. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, 2000.
- [2] V. Hatzivassiloglou and J. Wiebe. Effects of Adjective Orientation and Gradability on Sentence Subjectivity. In *Proceedings of the 18th conference on Computational linguistics*, Germany, 2000.
- [3] P. D. Turney. Thumbs up or thumbs down?: Semantic Orientation Applied to Unsupervised Classification of Reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, Philadelphia, Pennsylvania, USA, 2002.
- [4] N. Godbole, M. Srinivasaiah and S. Skiena. Large-scale Sentiment Analysis for News and Blogs. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, 2007.
- [5] S.-M. Kim and E. Hovy. Automatic Detection of Opinion Bearing Words and Sentences. In *Companion Volume to the Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)*, 2005.
- [6] E. Riloff and J. Wiebe. Learning Extraction Patterns for Subjective Expressions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2003.
- [7] J. Wiebe, T. Wilson and M. Bell. Identifying Collocations for Recognizing Opinions. In *Proceedings of the ACL Workshop on Collocation: Computational Extraction, Analysis, and Exploitation*, Toulouse, France, 2001.
- [8] Hong Yu and Vasileios Hatzivassiloglou. Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2003.
- [9] T. Wilson, J. Wiebe and R. Hwa. Just How Mad are You? Finding Strong and Weak Opinion Clauses. In *Proceedings of AAAI-04, 21st Conference of the American Association for Artificial Intelligence*, 2004.
- [10] Z. Fei, X. Huang and L. Wu. Mining the Relation between Sentiment Expression and Target Using Dependency of Words. In *proceedings of 20th Pacific Asia Conference on Language, Information and Computation (PACLIC20)*, Wuhan, China, 2006.
- [11] J. Yi, T. Nasukawa, R. Bunescu and W. Niblack. Sentiment Analyzer: Extracting Sentiments about a Given Topic using Natural Language Processing Techniques. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, 2003.
- [12] J. Yi and W. Niblack. Sentiment Mining in WebFountain. In *Proceedings of the International Conference on Data Engineering (ICDE)*, 2005.
- [13] M. Hu and B. Liu. Mining Opinion Features in Customer Reviews. In *Proceedings of AAAI*, San Jose, USA, 2004.
- [14] A.-M. Popescu, and O. Etzioni. Extracting Product Features and Opinions from Reviews". In *Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, Vancouver, British Columbia, Canada, 2005.
- [15] R. Ghani, K. Probst, Y. Liu, M. Krema and A. Fano. Text Mining for Product Attribute Extraction. SIGKDD Explorations Newsletter, 8(1). 2006.
- [16] A. Qadir. Identifying Frequent Word Associations for Extracting Specific Product Features from Customer Reviews. In *Proceedings of International Symposium on Data and Sense Mining Machine Translation and Controlled Languages, and their Application to Emergencies and Safety Critical Domains*, Besancon, France, 2009.
- [17] M.-C. de Marneffe and C. D. Manning. The Stanford typed dependencies representation. In *COLING Workshop on Cross-framework and Cross-domain Parser Evaluation*, 2008.
- [18] M.-C. de Marneffe and C. D. Manning. Stanford Typed Dependencies Manual. Technical report, 2008.
- [19] L. Zhuang, F. Jing and X. Zhu. Movie Review Mining and Summarization. In *Proceedings of ACM Conference on Information and Knowledge Management (CIKM)*, Arlington, Virginia, USA, 2006.
- [20] T. H. King, R. Crouch, S. Riezler, M. Dalrymple and R. Kaplan. The PARC 700 Dependency Bank. In *4th International Workshop on Linguistically Interpreted Corpora (LINC-03)*, 2003.



# Author Index

Balahur, Alexandra, 23

Boldrini, Ester, 23

Chen, Zheng, 17

Haralick, Robert, 17

Ji, Heng, 17

Jurgens, David, 9

Kosseim, Leila, 1

Lloret, Elena, 23

Margova, Ruslana, 32

Martínez-Barco, Patricio, 23

Mithun, Shamima, 1

Montoyo, Andrés, 23

Palomar, Manuel, 23

Qadir, Ashequl, 38

Stevens, Keith, 9

Temnikova, Irina, 32