Catching the news: two key cases from today

Ruslana Margova,
Geomedia magazine
naruslana@yahoo.com

Irina Temnikova,
University of Wolverhampton,
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 phemomena 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) and social networks like FaceBook (www.facebook.com) and Twitter (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), 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 then 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 occurencies 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-occurences.

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:

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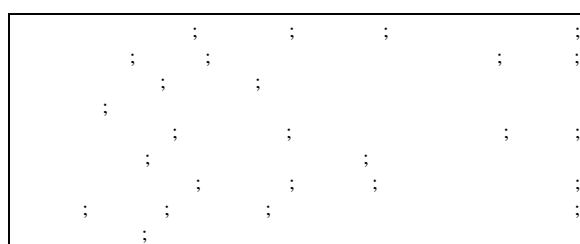
Another example is the list of proper names and co-references which do not appear in the news texts but are used by the bloggers:

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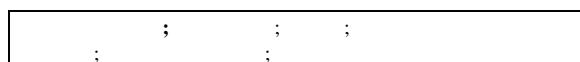
A further in-depth linguistic analysis may help the development of rule-based Information extraction of comments. Information extraction would improve the general knowledge about these people and would facilitate the better understanding of the problem.

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 26th, 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, *Palimsestes. 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 Benhjamins 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, Semeiotike Recherche pour une Semanalyse, 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 Christan 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/

[14] http://paulocoelhoblog.com/?s=iran

[15] http://www.ap.org/

[16] http://news.bbc.co.uk/