Facilitating online discussions by automatic summarization

Sander Wubben [a] Suzan Verberne [b] Emiel Krahmer [a]
Antal van den Bosch[b]

[a] *Tilburg Centre for Cognition and Communication*
Tilburg University
[b] *Centre for Language Studies/Centre for Language and Speech Technology*
Radboud University

Abstract

In the DISCOSUMO project, we aim to develop a computational toolkit to automatically summarize discussion forum threads. In this paper, we present the initial design of the toolkit, the data that we work with and the challenges we face. Discussion threads on a single topic can easily consist of hundreds or even thousands of individual contributions, with no obvious way to gain a quick overview of what kind of information is contained within the thread. We address the summarization of forum threads with domain-independent and language-independent methodology. We evaluate our system on data from four different web forums, covering different domains, languages and user communities. Our approach is largely unsupervised, using recurrent neural networks. Evaluation of the first version should point out where in the pipeline supervised techniques and/or heuristics are required to improve our summarization toolbox. If successful, the automatic summarization of discussion forum threads will play an important role in facilitating easy participation in online discussions.

1 Introduction

The way information is being consumed has changed drastically in recent years, due to, for example, the introduction of mobile devices and the rise of social media in general. We may still read news articles, but increasingly also find our information in other internet sources, including discussion fora and social media. News production has also shifted to a more participatory journalistic focus in addition to traditional news outlets. With the ubiquity of smartphones, stories can be brought live from the scene by civilians while they are happening. The shootings in Virginia Tech[1] and in the Aurora movie theater[2] were covered live on reddit forums by people present on the scene.

Interestingly, such sources are rather different from more traditional news texts, and are characterized by informal, subjective, and unedited language use. They also have a specific structure of time-stamped and threaded messages. Discussion threads on a single topic can easily consist of hundreds or even thousands of individual contributions, with no obvious way to gain a quick overview of what kind of information is contained within the thread. In practice, users currently need to read through the posts of the thread, both relevant and irrelevant ones, until they find what they are looking for.

The goal of the DISCOSUMO project that we introduce in this paper is to develop a toolbox for the automatic summarization of forum threads. The toolbox is meant to be modular, domain-independent, and language-independent where possible. In this paper we present the design of our system and illustrate the challenges of our project with the use of four discussion forums: two personal interest forums and two news forums.

Facilitating online discussions by making them more accessible is something we believe is a valuable societal issue in the digital age. Important information contained in these threads is made more accessible and as a result we hope that these technologies can help improve the quality of online discussions.

[1] http://newsfeed.time.com/2011/12/10/student-posts-live-reddit-qa-during-virginia-tech-lockdown/

[2] http://www.poynter.org/news/mediawire/181840/reddit-covers-the-aurora-movie-theater-shooting-dark-knight-rises/

2 Related work

In the field of natural language processing, automatic summarization is a well researched topic [7, 14]. Automatic summarization can be defined as the production of a shortened version of a document or set of documents that retains the most important information. Automatic summarization of documents has been shown to improve the speed of decision making for the reader, while not negatively impacting the decision[9]. Additionally, multi-document summarization may help reduce the time needed to find relevant information. This is particularly helpful in open-ended high-recall search tasks, where a user is searching for as much information about a topic as possible [12, 8].

Although automatic summarization is a lively area of research, the field is arguably progressing slowly. The current state-of-the-art automatic summarization only works well in specific domains such as scientific articles, news articles, e-mail messages, advertisements, and blogs, where the most important information tends to be located in predictable places [12]. These methods do not work well on texts in which the information is unpredictably spread throughout the text, as we find in internet forums [19].

Even though methods for automatic summarization are evolving, the features used in these systems tend to stay the same (word frequency, part of speech, position in the text) and usually extensive feature engineering and supervised learning is involved, making the system dependent of human-labeled training data, less robust and thus less suitable for wide adoption [7]. Our goal is to build a robust system that automatically learns a representation of the data, and relies on this representation for summarization. Metadata such as author profiles, search queries, and manually assigned post scores can be added as information for improving summarization.

A general characteristic of many automatic summarization systems is that they are *extractive*: the output summaries they produce is a selection of sentences that occurred verbatim in the input documents. An extractive approach is unlikely to yield satisfactory results when applied to discussion threads, for instance because the result can be expected to be incoherent (due to missing antecedents for pronouns in extracted sentences). The approach we envisage is *abstractive*: it will rely on phrases from the forum thread, but these will be modified and combined into newly generated sentences. This task, sentence modification, will rely on existing techniques for sentence paraphrasing, sentence compression and sentence fusion (e.g., [1, 10, 21, 11]). The generation of new sentences is a topic addressed in the field of Natural Language Generation (e.g., [15]).

Recently, work on text generation based on neural networks has seen a lot of potential in tasks such as sentiment analysis [16], image caption generation[18], machine translation [17] and question answering [4]. These approaches make use of *recurrent neural networks*: neural networks that can operate on sequences of data and feed their own output back into the network. This makes these networks suitable for sequential data such as text. An advantage of this approach is that the models trained in this way tend to learn a representation of the data by themselves, and can therefore be applied in various scenarios with little modification. We conjecture that this approach will be beneficial for the automatic summarization of discussion forums as well.

3 System design: Our approach

The system that we will develop takes forum threads as input and produces an abstractive summary as output. The summary should contain the important information in the thread and should be readable on mobile devices. It can well be the case that multiple opinions are expressed in a thread. The resulting summary should then show the summarized important opinions for the specific thread. The system should be highly modular. Figure 1 gives an indication of the system architecture.

The first step in the pipeline is to extract and preprocess the textual content from the forum thread. Additionally, information about the parent post, and the references between other posts in the thread should be extracted in order to preserve the hierarchical tree-like structure present in many forum threads. We process the data and feed the results into the *embedding model* . The embedding model generates a semantic representation for each post as follows: First, *word embeddings* are generated. Word embeddings capture the semantic meaning of a word by taking into account the context it appears in. So, the learned embedding for ‘journalist’ would be similar to the embedding for ‘reporter’, as both words tend to occur in similar contexts [13]. These word embeddings are then fed into the forum post encoder which learns to encode forum posts into a high-dimensional vector (the post embedding). We

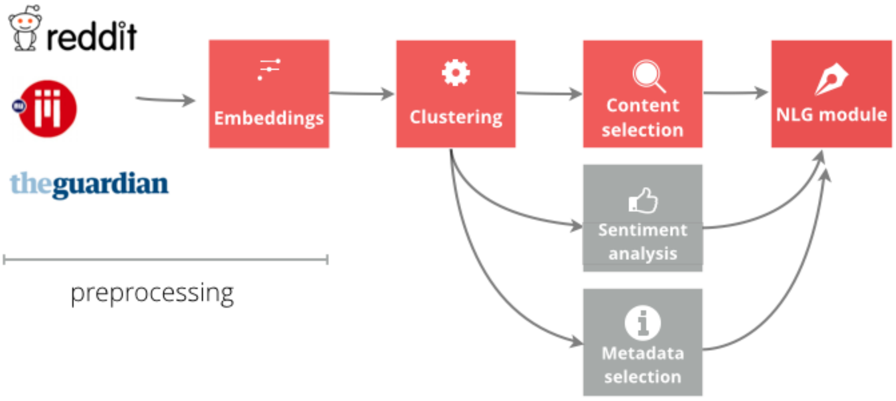

Figure 1: Architecture of the summarization toolkit. The system is modular: extra features or functionality can be added.

use an autoencoder architecture similar to the one described by [17]. The autoencoder learns a fixed low-dimensional vector representation for each forum post (that is, each post is represented by a fixed sequence of numbers).

Once we have these representations, we can feed them into the *clustering* module (See Figure 1). This module groups similar posts by calculating the similarity between the post vectors. The rationale behind this approach is that we expect clusters to form around the different opinions, but that they will also group off-topic parts of the thread.

The *content selection* component determines which information is sufficiently important to retain and filters out redundant and/or noisy information. We investigate several statistical and machine-learning-based methods for filtering, ranking and clustering the information. Note that this is a core component of the system as the quality of the selected information largely determines the overall quality of the output summary. Besides content extraction, we also extract statistical information of the discussion threads, such as the percentage of posts that express the same viewpoint and percentage of off-topic posts in the thread.

Additionally, *sentiment analysis* is applied to extract opinions in the form of sentiment labels. User-contributed labels such as ratings or likes of posts may be used directly as training labels [6] by which a machine-learning classifier can be trained.

The final step of *text generation* in the system architecture aims to produce a readable, fluent summary of information and opinions in the source documents. We apply and evaluate different types of text-to-text generation techniques, such as sentence simplification [21], sentence fusion [5] and paraphrasing [22]. Additionally, we employ natural language generation methods to transform the forum thread statistics and meta data to readable text [3].

Models will be developed to adapt the generated summary text to different devices. A summary read on a laptop could be longer than one read on a tablet or smartphone. Additionally, personalization can be applied to tailor the text to the goals and intents of the specific user. For readers who enter a forum thread through a search query, we use the content of the query to adapt the summary to meet the reader's intent, a task called *query-based summarization* [2].

While each step will be evaluated thoroughly, the final evaluation will consist of actual implementation of the models in the internal Sanoma Media[3] BV test architecture and additionally on their websites. User statistics and evaluations will be collected and analyzed in order to evaluate and improve the summarization models. In particular, we will study whether the availability of summarized discussion threads facilitates the contribution to discussions from mobile devices and may lead to improved discussions themselves as it becomes easier for users to quickly grasp the contents of a discussion.

[3] Sanoma Media is a Dutch company, part of the Finnish media concern Sanoma. Sanoma Media is the publisher of dozens of (paper and online) titles, among which Nujij and Viva.

Table 1: The four discussion forums used for the development and evaluation of our system. Note that personal interest does also cover news topics, as the examples of reddit forum posts in Section 1 illustrate.

Forum name	URL	Type	Language	Upvotes/downvotes?	# registered users
The Guardian	www.guardian.co.uk	News	English	Upvotes only	Unknown
Nujij	www.nujij.nl	News	Dutch	Both	90K
Reddit	www.reddit.com	Personal Interest	English	Both	3.5M
Viva forum	forum.viva.nl	Personal Interest	Dutch	None	1.5M

4 Data

We will evaluate our system on data from four different web forums. In the next subsection we will briefly describe them. In Section 4.2, we present a generalized XML format that we use for processing the forum data.

4.1 Web forums

Table 1 gives an overview of the four discussion forums that we use for the development and evaluation of our summarization system.

The Guardian is a British newspaper. The news articles on the website have been written by journalists or columnists and have been professionally edited. Visitors of the Guardian web site can post comments below news articles if they have registered as user. The comments below a news article are ordered chronologically, with the newest comment first. Users can reply to other users, thereby creating a hierarchy of posts, and may give upvotes to other users' posts.

Nujij is a discussion forum connected to the online Dutch newspaper nu.nl. An important difference with the Guardian is that on Nujij, every user can start a thread. The opening post is always a link to a news article or blog post, together with a brief summary and a picture. Users can vote for articles, and give upvotes and downvotes to other users' comments. Comments can contain references to one or more previous comments. Nujij has 90.000 registered users, approximately 200 threads are started each day, and the number of visits per month is 5.935.000 (737.000 unique visitors).[4]

reddit is an online forum where any registered user can start threads in the form of messages and links. Users can post comments, reply to other users' posts (creating hierarchically structured threads) and give upvotes and downvotes to other users' posts. In June 2015, reddit had 164 Million visitors, of which 3,5 Million registered users. We obtained reddit's entire publicly available comment dataset, 1.7 billion JSON objects "complete with the comment, score, author, subreddit, position in comment tree and other fields that are available through Reddit's API"[5].

Viva Forum is a Dutch web forum with a predominantly female user community. Registered users can start new threads and comment on threads. Many opening posts contain questions asking advice but there are also threads for life experience sharing. Threads do not contain any explicit hierarchy, but users can use the quote option to directly respond to another user's post. Viva forum has 19 Million page views per month (1.5 Million unique visitors).[6]

4.2 Generalized XML format

We define an XML format for forum data in the form of a DTD (document type definition), which is shown in Figure 2. The root of the XML structure is a thread. A thread consists of one or more posts, the first post being the opening post.

The hierarchy of the thread is realized by including an element 'parent' for each post. The opening post does not have a parent; each next post has 1 or more parents: the post id(s) of the post(s) it replies

[4] http://www.sanoma.nl/merken/bereik/nujijnl/

[5] Downloaded from http://www.reddit.com/r/datasets/comments/3bxlg7/i_have_every_publicly_available_reddit_comment/ on July 15, 2015

[6] http://www.sanoma.nl/merken/bereik/viva/

```
<!ELEMENT thread (threadid,post+,category*,type*,nrofviews?)>
<!ELEMENT post (postid,author,timestamp,
                            parent*,upvotes?,downvotes?,body)>
<!ELEMENT author (#PCDATA)>
<!ELEMENT timestamp (#PCDATA)>
<!ELEMENT parent (#PCDATA)>
<!ELEMENT upvote (#PCDATA)>
<!ELEMENT downvote (#PCDATA)>
<!ELEMENT body (content,url*)>
<!ELEMENT content (#PCDATA)>
<!ELEMENT url (#PCDATA)>
```

Figure 2: Document type definition (DTD) for forum threads. An asterisk (*) means zero or more occurrences; a plus (+) means one or more occurrences; a question mark (?) means zero or one occurrences. Category and type can be used for forum-specific purposes, such as the name of the subforum the thread belongs to, or a classification label added in our analysis. Upvotes and downvotes are counts.

to. The thread structure of reddit is truly hierarchical (a depth of 10 replies is no exception). The thread structure of the Guardian is only hierarchical up to one level (if comment B is a reply to comment A, then comment B cannot receive replies itself). Viva forum threads are structurally flat: all comments reply to the opening post, although it is possible to refer to previous posts using quoteblocks. In Nujij threads, a comment can refer to more than one previous posts, using '@#' in the text, in which # is the index number of the post it refers to.

5 Challenges

The general challenge we are facing is to make our toolbox domain-independent and language-independent where possible. As introduced in Section 3, our approach is largely unsupervised (it doesn't require labeled data), using word embeddings in a recurrent neural network. Evaluation of the first version should point out where in the pipeline supervised techniques and/or heuristics are required to improve our summarization toolbox.

5.1 Distinguishing relevant from non-relevant posts

In a pilot experiment, we manually labeled a sample of 10 threads from the Viva forum. Two annotators were asked to label the posts with either YES or NO. According to these annotations, about 60% of posts is relevant. However, there are large differences between threads: for some threads 80% of the posts is relevant, while for other threads, 80% is irrelevant.

The content selection module in our system will take care of this element and careful evaluation and tuning is required to investigate to which degree the selections of this module will correspond with human judgements.

5.2 Recognizing the type and structure of the thread

Intuitively, the type of forum (news, personal interest) and the type of thread (question, opinion poll, life experience sharing) should determine the summarization strategy. For example:

- If a closed question is asked (e.g. "ms office to libre/open office?" on reddit), counting votes and presenting the counts with the most important arguments would make an informative summary.
- If the topic is more complex (e.g. the Guardian article "Turkey says Kurdish peace process impossible as Nato meets"), counting votes is not sensible and a summary of the different points of views will probably be the most useful to the reader.
- If the topic is very personal (e.g. "I'm 28 years old, I have breast cancer. Here's my story." on reddit), then summarizing the thread may not be helpful at all for the interested reader.

This shows that before starting summarization, we need categorization of the thread. The number and granularity of categories, and the consequence for the summarization strategy are still open questions.

Apart from the thread category, it is necessary to recognize the structure of a thread. The first step in extracting the thread structure is identifying the parent of each post. In the reddit data, the parent id is given for each post. In creating the XML structure for Nujij and Viva, we need to use the formatting of quoteblocks and/or references in the text to find the parent of each post.

In previous work [20], the structure of a thread is defined much more elaborately than by recognizing the parent of each post. The authors propose a classification scheme for defining the *type of interaction* between a post and its reply. They distinguish 12 interaction categories among which 'additional question', 'confirmation', 'answer', and 'correction'.

In our pilot study, we let participants use this scheme to classify posts from three Viva forum threads. We found that agreement was generally very low. Therefore, we prefer not to use an extensive classification scheme at this point. After development and evaluation of the first version of our system, we will be able to judge the need for a classification scheme defining thread structure.

5.3 Personalization/diversification

Not all readers are interested in the same aspects of a discussion. If a reader enters the discussion through a search query, we have some information about his interest in the discussion. We can employ this information to create a tailored summary, focusing on specific aspects of the discussion. For example, a visitor searching for opinions on the republican candidate Romney, may enter the reddit thread "Which potential Republican candidate for President in 2016 stands the best chance of winning in the general election". Another visitor searching for opinions on Jeb Bush may enter the same thread. They both require a different summary: the first about Romney, the second about Bush. To solve this challenge, we will adapt techniques from the field of query-based summarization [2] to the specifics of web forum data.

6 Conclusion

In this paper we introduced the DISCOSUMO project. In this project we will develop a computational toolkit to automatically summarize subjective information in online discussion threads. We identified several important issues, such as automatically detecting relevant posts, clustering viewpoints and representing them in the generated summary. We expect that the unique subjective perspective on news stories brought by online resources will continue to play an important role in the media landscape and we aim to facilitate and improve production and consumption of news stories through these channels by generating informative summaries using state of the art computational methods.In this manner, relevant information should be easier to find and analyze, improving the accessibility and quality of online discussions.

References

[1] Regina Barzilay and Kathleen R McKeown. Sentence fusion for multidocument news summarization. *Computational Linguistics*, 31(3):297–328, 2005.

[2] Olga Feiguina and Guy Lapalme. Query-based summarization of customer reviews. In *Advances in Artificial Intelligence*, pages 452–463. Springer, 2007.

[3] Albert Gatt and Ehud Reiter. Simplenlg: A realisation engine for practical applications. In *Proceedings of the 12th European Workshop on Natural Language Generation*, pages 90–93. Association for Computational Linguistics, 2009.

[4] Mohit Iyyer, Jordan Boyd-Graber, Leonardo Claudino, Richard Socher, and Hal Daumé III. A neural network for factoid question answering over paragraphs. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 633–644, 2014.

[5] Emiel Krahmer, Erwin Marsi, and Paul van Pelt. Query-based sentence fusion is better defined and leads to more preferred results than generic sentence fusion. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pages 193–196. Association for Computational Linguistics, 2008.

[6] Florian Kunneman, Christine Liebrecht, and Antal van den Bosch. The (un)predictability of emotional hashtags in Twitter. In *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)*, pages 26–34, Gothenburg, Sweden, April 2014. Association for Computational Linguistics.

[7] Elena Lloret and Manuel Palomar. Text summarisation in progress: a literature review. *Artificial Intelligence Review*, 37(1):1–41, 2012.

[8] Manuel J Maña-López, Manuel De Buenaga, and José M Gómez-Hidalgo. Multidocument summarization: An added value to clustering in interactive retrieval. *ACM Transactions on Information Systems (TOIS)*, 22(2):215–241, 2004.

[9] Inderjeet Mani, Gary Klein, David House, Lynette Hirschman, Therese Firmin, and Beth Sundheim. Summac: a text summarization evaluation. *Natural Language Engineering*, 8(01):43–68, 2002.

[10] Erwin Marsi and Emiel Krahmer. Classification of semantic relations by humans and machines. In *Proceedings of the ACL workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 1–6. Association for Computational Linguistics, 2005.

[11] Erwin Marsi, Emiel Krahmer, Iris Hendrickx, and Walter Daelemans. On the limits of sentence compression by deletion. In *Empirical methods in natural language generation*, pages 45–66. Springer, 2010.

[12] Kathleen McKeown, Rebecca J Passonneau, David K Elson, Ani Nenkova, and Julia Hirschberg. Do summaries help? In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 210–217. ACM, 2005.

[13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[14] Ani Nenkova, Sameer Maskey, and Yang Liu. Automatic summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts of ACL 2011*, page 3. Association for Computational Linguistics, 2011.

[15] Ehud Reiter, Robert Dale, and Zhiwei Feng. *Building natural language generation systems*, volume 33. MIT Press, 2000.

[16] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

[17] Ilya Sutskever, Oriol Vinyals, and Quoc VV Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[18] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. *arXiv preprint arXiv:1411.4555*, 2014.

[19] Nayer Wanas, Motaz El-Saban, Heba Ashour, and Waleed Ammar. Automatic scoring of online discussion posts. In *Proceedings of the 2nd ACM Workshop on information Credibility on the Web*, pages 19–26. ACM, 2008.

[20] Li Wang, Su Nam Kim, and Timothy Baldwin. Thread-level analysis over technical user forum data. In *Australasian Language Technology Association Workshop 2010*, pages 27–31, 2010.

[21] Sander Wubben, Antal Van Den Bosch, and Emiel Krahmer. Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 1015–1024. Association for Computational Linguistics, 2012.

[22] Sander Wubben, Antal Van Den Bosch, and Emiel Krahmer. Creating and using large monolingual parallel corpora for sentential paraphrase generation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may 2014. European Language Resources Association (ELRA).