Feature Selection for Language Independent Text Forum Summarization

Vladislav A. Grozin(✉), Natalia F. Gusarova, and Natalia V. Dobrenko

National Research University of Information Technologies, Mechanics and Optics,
Saint-Petersburg 197101, Russia
grozin@my.ifmo.ru

Abstract. Nowadays the need for multilingual information retrieval for searching relevant information is rising steadily. Specialized text-based forums on the Web are a valuable source of such information. However, extraction of informative messages is often hindered by large amount of non-informative posts (the so-called *offtopic* posts) and informal language commonly used on forums.

The paper deals with the task of automatic identification of posts potentially useful for sharing professional experience within text forums irrespective of the forum's language. For our experiments we have selected subsets from various text forums containing different languages. Manual markup was held by native speaking experts. Textual, thread-based, and social graph features were extracted. In order to select satisfactory language-independent forum features we used gradient boosting models, relative influence metric for model analysis, and NDCG metric for measuring selection method quality.

We have formed a satisfactory set of forum features indicating the post's utility which do not demand sophisticated linguistic analysis and is suitable for practical use.

1 Introduction

Nowadays we are facing the rapid growth of non-English documents on the Internet. The need for multilingual information retrieval and language-independent information access for professionals, organizations and businesses is rising steadily. Specialized text forums are a valuable source of knowledge of that kind. Forums contain experience of people who actually used the technology and its features, often expressed in their native language. Forums contain both positive and negative experience—something that is not available from official documentation at all. But they also contain a lot of trivial, repeated and still irrelevant posts. Therefore the expert not knowing forum language should have opportunity to extract useful and informative posts in order to study them in more detail subsequently.

The obvious solution is to use techniques of text summarization. But important information can be provided in different languages, including highly inflected, having complex grammar and rather weak text analysis tools. It is

P. Klinov and D. Mouromtsev (Eds.): KESW 2015, CCIS 518, pp. 63–71, 2015.
DOI: 10.1007/978-3-319-24543-0_5

a challenge for using parsers, part-of-speech taggers, morphological analyzers and full dictionaries for any of the languages. In fact, most application for text processing are monolingual tools or tools covering a few commonly spoken languages. In [17] the lack of linguistic resources is called "one obvious bottleneck for the development of multilingual tools". So, the procedure of text forum summarization has to be simple and language-independent in order to be used in practice.

In this paper, we address the task of automatically identifying posts potentially useful for sharing professional experience within text forums irrespective of forum language. We aim to choose a reasonable set of forum features indicating the post's utility which doesn't demand sophisticated linguistic analysis and is suitable for practical use.

2 Related Work

The task of Web Forum Thread Summarization typically aims to give a brief statement of each thread that involving multiple dynamic topics. Traditional summarization methods are cramped here by some challenges [15]. The first is topic drifting: as the post conversation progresses, the semantic divergence among subtopics will be widened. Besides, most posts are composed of short and elliptical messages, their language is highly informal and noisy, and traditional text representation methods have sufficient limitations here.

According to the survey in [15], the majority of works in the area of forum summarization use extraction-based techniques [16] and single-document approach. A lot of research on automatic dialogue summarization use corpus-based and knowledge-based methods. For example, authors [23] identify clusters in the Internet relay chats and then employ lexical and structural features to summarize each cluster. Authors [15] have proposed a forum summarization algorithm that models the reply structures in a discussion thread. In order to represent information of online forum in a learning environment author [5] uses concept-based summarization: each word in the document is labeled as a part of speech in grammar, and to handle the word sense disambiguation problem similarity measures based on WordNet is used. Statistical methods of dialogue summarization are also of great interest. For example, in [20] unsupervised (TF-IDF and LDA topic modeling) and supervised clustering procedures (using SVMs and MaxEnt) are used in combination for decision summarization for spoken meetings. Authors [4] consider the problem of extracting relevant posts from a discussion thread as a binary classification problem where the task is to classify a given post as either belonging to the summary or not. In general statistical methods are very various, including genetic algorithms [6], hybrid approaches [13, 18], an integer linear programming approach [2], and so on.

There is a number of the works devoted to multi-lingual aspects of text summarization. For example, in order to fulfill sentiment analysis of multi-lingual Web resource [12] consider English as basic and use language-specific semantic lexicons of sentiment-carrying words. Contrary to this approach, authors [3] show that the multilingual model consistently outperforms the cross-lingual one. Practical experience of developing natural language processing applications for many languages is described in [17]. The author considers Machine Learning methods as an extremely promising approach to develop highly multilingual systems.

A fast-growing number of studies have shown that the social factor can be useful in text forum summarization. For example, authors [14] apply similar measures as used in blogs to the forums, such as counting the number of common tags and replying or citing the same threads. Authors [22] explain that in an online forum context a central core (strongly connected component) contains users that frequently help each other by following questioner (requester) answerer (expert) links.

Feature categories being used in forum analysis studies mainly depend on specifics of a task. For example, authors [1] are concerned with classifying sentiments in extremist group forums. To do this they use syntactic, semantic, link-based, and stylistic features. In [4] thread-specific features are used for identifying subjectivity orientation of threads. These include structural features, dialog act features, subjectivity lexicon based features and sentiment features. Authors [10] use textual features for detecting the reputation dimension of a post. Authors [8,19] extract contexts and answers of questions from online forums using discourse and lexical features as well as non-textual and structural features. But in general the selected features are highly language-dependent and need complicated techniques for their analysis.

To sum up, we can say that information retrieval within text forums irrespective of its language represents a complex problem, and its decision in an explicit form isn't submitted in literature. Methods of dialogue summarization based on machine learning algorithms showed good prospects, but there is a great need of simple and language-independent techniques.

3 Methods

For our experiments we have selected some forums held in highly inflective languages with complex grammar and rather weak text analysis tools, in particular - German, Russian and Chinese (Mandarin). Also, we selected one forum held in English for comparison. Detailed information about the forums is presented in Table 1. Within each forum we selected thread of interest. Each posts' usefulness was manually marked down by experts (Table 2). We have invited experts of the relevant field who are native speakers in the languages of the forums.

As mentioned above, there is a lot of works proposing different features for text forums, potentially suitable for usefulness evaluation. However, not all of them are suitable for machine learning due to the specifics of our task since we need language-independent features. We have chosen features applicable for multilinguistic approach. The list of the chosen features is presented in Table 3.

We calculated text sentiment value using stemmed sentiment keywords and word parts specific for the forums language as well as stemming technique. The

Table 1. The chosen Internet forums, their language, topics, keywords, and statistics

Forum	Language	Topic	Threads/posts	Keywords
1 gamedev.ru	Russian	Unity	10/410	unity
2 hifi-forum.de	German	Windows vs Linux	13/173	windows, linux
3 forum.modelsworld.ru	Russian	Ship modeling	3/150	ship, model
4 5500.forumactif.org	French	Ship modeling	3/150	ship, model
5 bbs.csdn.net	Chinese	cocos2d-x	11/120	cocos
6 bbs.chinaunix.net	Chinese	Linix for beginners	11/103	linux
7 knittinghelp.com	English	Knitting techniques	43/450	knit

Table 2. Thread's usefulness scale

Value	Comment
0	Offtopic
1	Post is on the chosen topic, but argumentation is incomplete or absent
2	Post is on the chosen topic, and the authors point of view is well-justified with explanations or external links

resulting values were normalized to the range from -1 (strongly negative text) to +1 (strongly positive text).

Also, simple non-semantic text features were extracted: text length in characters, number of links and number of keywords. Keywords were chosen strictly corresponding to the name of the forum topic. A more extensive list of keywords would mean a search for synonyms and equivalents, but it requires semantic analysis.

We represented social structure in the form of a social graph, where the nodes are the users, and edges indicate a link between two users. For the creation of the social graph we have used citation analysis: if person A quotes person B by explicitly mentioning his name in text, there is a guaranteed connection between A and B. We used two methods: a non-sentiment graph (edge weight is always 1) and a sentiment graph (edge weight is related to the posts sentiment value). After the creation of the graph parallel edges weights were summed. Then, the weights of the edges were inverted.

Node centrality is often used to find people who are important members of society. We considered some proven [7,21] metric to evaluate node centrality: Betweenness centrality - the number of shortest paths between all pairs of nodes that pass through the node; inDegree - the total weight of incoming edges; outDegree - the total weight of the outgoing edges.

Position in thread is calculated as number of post in chronological order (first post has position in thread equal to one, next one is equal to two etc.).

In order to select features indicating post usefulness we need models to capture dependence of usefulness on features. We used gradient boosting models in "gbm v.2.1" package (gbm) with following settings:

Table 3. Selected features

Type	Feature	What this feature means
Posts author graph features	Betweenness, non-sentiment graph	Authors social importance
	inDegree, non-sentiment graph	How many times author was quoted
	outDegree , non-sentiment graph	How many times author quoted someone
	Betweenness, sentiment graph	Authors social importance
	inDegree, sentiment graph	With which sentiment author was quoted
	outDegree, sentiment graph	Authors quotes sentiment
Posts author features	Number of threads author is participating in	Author activity
Thread-based post features	Position in thread	Chance of off-topic
	Times quoted	Posts impact on forum
Text features	Length	Number of arguments and length of explanations
	Links	Number of external sources/images
	Sentiment value (calculated using sentiment keywords)	Posts usefulness
	Number of keywords	Topic conformity

- CV folds: 3
- Shrinkage: 0.005
- Number of trees: 4000

Other parameters were left default. Models were constructed for each language independently.

In order to estimate quality of constructed models to ensure that models constructed from features listed in Table 3 give the same results for each language, i.e. are language-independent. To do this we divided available data into training set (60% of each forum) and test set (40%). Widely used recall/precision metrics are not applicable due to the fact that we have three Utility levels, and those metrics are used for estimating binary classification quality. To estimate quality of our models we follow these steps:

1. Fit model to train set.
2. Apply model to test set; it gives $\widetilde{Utility}$ - some approximation of true $Utility$ values of test set.
3. Sort posts in descending order by $\widetilde{Utility}$, and take N top posts. This gives selection of N best posts according to model.
4. Calculate NDCG metric:

$$NDCG_N = \frac{DCG_N}{IDCG_N}; DCG_N = rel_1 + \sum_{i=2}^{N} \frac{rel_i}{log_2(i)},$$

where N is number of selected posts, $rel(i)$ is quality (i.e. true $Utility$) of i-th selected post, and $IDCG_N$ is normalization coefficient calculated as maximum possible DCG_N for specified dataset and N. This metric lies between 0 and 1 (assuming non-negative $rel(i)$), and is cross-query comparable. It is commonly used for calculating ranking method quality.

In order to ensure stability of solution quality on different input sets we used bootstrap resampling-based method. Training set and test set for each forum were resampled with replacement before fitting model and calculating NDCG. This process was repeated 100 times. In each iteration some records were sampled out, and in fact in every iteration we had different training and test sets, so NDCG changed from iteration to iteration. Then, mean NDCG and 0.01 and 0.99 quantiles were calculated for each "language and N" pair, giving mean with confidence intervals for each language and N. 0.01 and 0.99 quantiles were also calculated for each N (ignoring language) to estimate overall confidence interval.

For baseline our experts formed an extensive list of keywords related to the topic of each chosen forum and their list of synonyms, i.e. semantic core (up to 50 words per forum). Then we applied stemming and lemmatization where possible (R package) to each post and these keywords, and counted amount of keywords and their stems in each post. After that we built linear regression model with this count and semantic value as features and post Utility as target variable. By doing this we emulated operation of a information retrieval system aware of forum language, its syntax and semantics and context of chosen narrow topic.

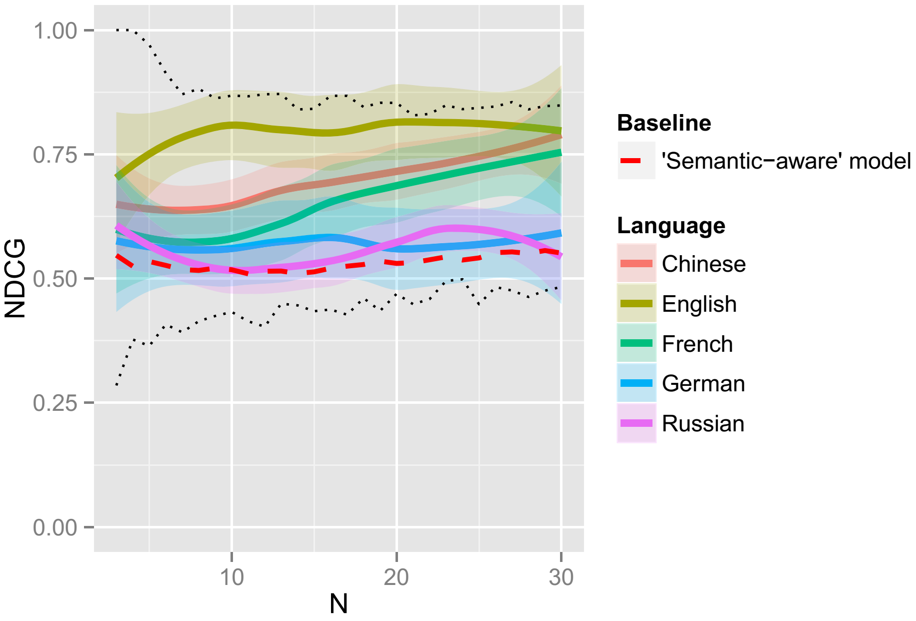

Fig. 1. Dependence of NDCG on language and size of selection

4 Results and Discussion

Fig. 1 shows the dependence of quality of selection (mean NDCG) with confidence intervals (0.01 and 0.99 quantiles) on the language and N. Our shortest forum contains 100 messages (and test set has 40), so we evaluated metrics for N varying from 1 to 30. Dotted lines represent 0.01 and 0.99 quantiles calculated for all available dataset (i.e. language is ignored). Also, simple baseline model performance (red dashed line) is drawn.

The analysis of dependence allows drawing the following conclusions:

- Allocation quality of our methods are generally better than baseline method. Also, baseline model requires knowledge extensive list of synonims and words related to chosen topics, forum language knowledge and complex forum post preprocessing. Our models are simple and stable, and do not require such things. It should be noted that our baseline uses sentiment value as a feature; as explained below, it is one of the most important features, so decent baseline performance is expected. Moreover, amount of semantic core words

Table 4. Features with the highest Relative Influence (RI)

Language	Top features ordered by relative influence
Chinese	Sentiment value
	Text length
	Position in thread
	Number of keywords
	Number of links
Russian	Sentiment value
	Text length
	Author betweenness, non-sentiment graph
	Number of keywords
	Position in thread
German	Text length
	Position in thread
	Sentiment value
	inDegree, non-sentiment graph
	outDegree, sentiment graph
French	Sentiment value
	Text length
	Number of threads the author participates in
	Number of keywords
English	Text length
	Sentiment value
	Author betweenness, non-sentiment graph
	outDegree, sentiment graph
	Number of keywords

in text (another feature in baseline model) is highly corellated with texts length, another important feature in out models.
- for French, German, Chinese and Russian confidence intervals overlap, and lines lie within each others confidence intervals ($NDCG = 0.63 \pm 0.07$). Better $NDCG$ values were received for English ($NDCG = 0.82 \pm 0.07$). It is apparently connected with overall simplicity of English language.

We also have to investigate which of selected features were the most important for each language. To do so we used Relative Influence (RI) metric [9] for each model. Top five features with the highest RI are presented in Table 4. Each column represents language, each Nth row is Nth top feature for each language. Features are sorted in descending order by their relative influence. Note that there are recurring top features: “Sentiment value”, “Text length”, “Position in thread”. Therefore it is reasonable to assume that those features are language-independent.

5 Conclusion

In this paper, we have addressed the task of automatically identifying posts potentially useful for sharing professional experience within technical text forums irrespective of forum language. We have shown that it is possible to allocate a reasonable set of forum features indicating the posts utility which doesn’t demand sophisticated linguistic analysis and is suitable for practical use. In our future work we plan to design more sophisticated models for feature selection, usage of complex features and considering quality of forum moderation.

References

1. Abbasi, A., Chen, H., Salem, A.: Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums. The University of Arizona (2007). http://ai.arizona.edu/intranet/papers/AhmedAbbasi_SentimentTOIS.pdf
2. Alguliev, R.M., Aliguliyev, R.M., Hajirahimova, M.S., Mehdiyev, C.A.: MCMR: Maximum coverage and minimum redundant text summarization model. Expert Systems with Applications **38**, 14514–14522 (2011)
3. Banea, C., Mihalcea, R., Wiebe, J.: Sense-level subjectivity in a multilingual setting. Computer Speech and Language **28**, 7–19 (2014)
4. Biyani, P., Bhati, S., Caragea, C., Mitra, P.: Using non-lexical features for identifying factual and opinionative threads in online forums. Knowledge-Based Systems **69**, 170–178 (2014)
5. Carbonaro, A.: WordNet-based Summarization to Enhance Learning Interaction Tutoring. Peer Reviewed Papers **6**(2) (2010)
6. Chen, J.-S., Hsieh, C.-L., Hsu, F.-C.: A study on Chinese word segmentation: Genetic algorithm approach. Information Management Research **2**(2), 27–44 (2000)
7. Ding, S.L., Cong, G., Lin, C.Y., Zhu, X.Y.: Using conditional random fields to extract contexts and answers of questions from online forums. In: Proceedings of the 46th Annual Meeting of the Association of Computational Linguistics, Columbus, Ohio, pp. 710–718. ACL (2008)

8. Freeman, L.C.: Centrality in social networks: Conceptual clarification. Social Networks **1**, 215–239 (1978)
9. Friedman, J.: Greedy boosting approximation: a gradient boosting machine. Ann. Stat. **29**, 1189–1232 (2001)
10. Garbacea, C., Tsagkias, M., de Rijke, M.: Feature Selection and Data Sampling Methods for Learning Reputation Dimensions. The University of Amsterdam at RepLab 2014 (2014). http://ceur-ws.org/Vol-1180/CLEF2014wn-Rep-GarbaceaEt2014.pdf
11. Generalized Boosted Regression Models. http://cran.r-project.org/web/packages/gbm/index.html
12. Hogenboom, A., Heerschop, B., Frasincar, F., Kaymak, U., de Jong, F.: Multilingual support for lexicon-based sentiment analysis guided by semantics. Decision Support Systems **62**, 43–53 (2014)
13. Huang, C.-C.: Automated knowledge transfer for Internet forum. Master thesis, Graduate School of Information Management, I-Shou University, Taiwan, ROC (2003)
14. Li, Y., Liao, T., Lai, C.: A social recommender mechanism for improving knowledge sharing in online forums. Information Processing and Management **48**, 978–994 (2012)
15. Ren, Z., Ma, J., Wang, S., Liu, Y.: Summarizing web forum threads based on a latent topic propagation process. In: CIKM 2011, October 24–28, Glasgow, Scotland, UK (2011)
16. Jones, K.S.: Automatic summarising: the state of the art. Information Processing and Management, Special Issue on Automatic Summarising (2007)
17. Steinberger, R.: Challenges and methods for multilingual text mining. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.167.4724
18. Tao, Y., Liu, S., Lin, C.: Summary of FAQs from a topical forum based on the native composition structure. Expert Systems with Applications **38**, 527–535 (2011)
19. Wang, B., Liu, B., Sun, C., Wang, X., Sun, L.: Thread Segmentation Based Answer Detection in Chinese Online Forums. Acta Automatica Sinica **39**(1) (2013)
20. Wang, L., Cardie, C.: Summarizing decisions in spoken meetings. In: Proceedings of the Workshop on Automatic Summarization for Different Genres, Media, and Languages, Portland, Oregon, June 23, 2011, pp. 16–24. Association for Computational Linguistics (2011)
21. White, D.R., Borgatti, S.P.: Betweenness centrality measures for directed graphs. Social Networks **16**, 335–346 (1994)
22. Yang, S.J.H., Chen, I.Y.L.: A social network-based system for supporting interactive collaboration in knowledge sharing over peer-to-peer network. International Journal of Human Computer Studies **66**(1), 36–40 (2008)
23. Zhou, L., Hovy, E.: Digesting virtual geek culture: the summarization of technical internet relay chats. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, ACL 2005, Stroudsburg, PA, USA, pp. 298–305. Association for Computational Linguistics (2005)