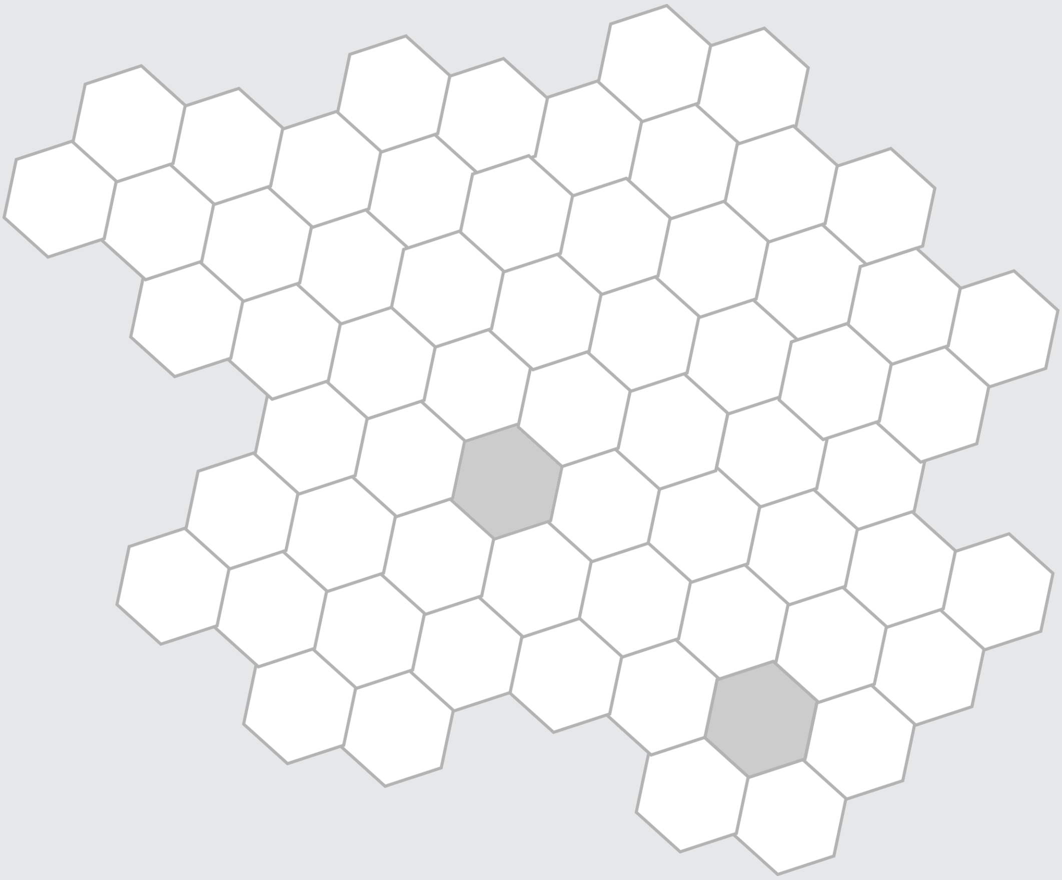
International Journal on SOCIAL MEDIA

MMM: Monitoring, Measurement, and Mining



II, 2011, 1–2
ISSN 1804-5251

International Journal on SOCIAL MEDIA

MMM: Monitoring, Measurement, and Mining

II, 2011, 1–2

Editor-in-Chief: Jan Žižka

Publisher's website: `www.konvoj.cz`
E-mail: `konvoj@konvoj.cz`, `SoNet.RC@gmail.com`
ISSN 1804-5251

Journal Profile

International Journal on Social Media MMM: Monitoring, Measurement, and Mining is an international scientific and refereed journal focused on questions and progress of social media, especially on their monitoring, measurement, analysis, and mining in social networks, e.g. Sentiment/Opinion Analysis in Natural-Language Text Documents, Algorithms, Methods, and Technologies for Building and Analysing Social Networks, Applications in the Area of Social Activities, Knowledge Mining and Discovery in Natural Languages Used in Social Networks, Medical, Economic, and Environmental Applications in Social Networks, etc.

International Journal on Social Media MMM: Monitoring, Measurement, and Mining seeks to share new knowledge, processes and methods. The journal publishes original works, project solutions, case studies, reviews and educational papers. Submitted papers should not have been previously published nor be currently under consideration for publication elsewhere.

Two issues per year provide a forum for distinguished as well as young authors.
A shortened thesis as well as final reports of projects supported by grant agencies are accepted for publishing.

International Journal on Social Media MMM: Monitoring, Measurement, and Mining is published with the partial financial support of Autonomous University of Barcelona and University of Wolverhampton.

All papers are refereed through a peer review process.

Submissions should be send in the PDF form via email to the following address: `SoNet.RC@gmail.com`.

Accepted papers are to be prepared according to the instructions available at `http://www.konvoj.cz/journals/mmm/`.

International Journal on Social Media MMM:
Monitoring, Measurement, and Mining,
II, 2011, 1–2
Published by KONVOJ, spol. s r. o. (ID CZ 47915391),
Bystřinova 4, CZ 612 00 Brno

The journal comes out twice a year. Printed in the Czech Republic.

Registration number of Ministry of Culture: MK ČR 19640
Issued in Brno, December 30, 2011.

ISSN 1804-5251

Contents

Editorial

Dear Readers,

in this issue of the International Journal Social Media MMM: Monitoring, Measurement, and Mining, we would like to present five interesting articles that are demonstrating several modern and interesting topics connected to automatic discovery of opinion or sentiment that are hidden within large volumes of data represented by natural languages. As you all certainly know very well, this area is now intensively studied because this kind of data extremely quickly increases from the volume point of view and, in addition, covers tens of natural languages from all corners of our more and more globalised world. Results of such research works are undoubtedly quite attractive in many different areas because knowing the sense and direction of the stored opinions can help improve our world.

Kristýna Machová and Tomáš Rakušinec deal with social networks, their types, analysis of their behaviour and properties. They focus on the dynamic analysis of social networks in the article 'Social Web Mapping Dynamic Analysis of Social Networks'.

As we all know, we can often meet with various emotions when communicating with our friends, colleagues, or unknown people. What can we do with emotions that are incorporated into textual data? A reader can find some aspects in the article 'Development of Japanese WordNet Affect for Analysing Emotions in Text' submitted by the authors Yoshimitsu Torii, Dipankar Das, Sivaji Bandyopadhyay, and Manabu Okumura.

Because the source of data are people with their typically subjective approaches, this problem should also be carefully analysed. The research work provided by Muhammad Abdul-Mageed and Mona Diab in their contribution 'Linguistically-Motivated Subjectivity and Sentiment Annotation and Tagging of Modern Standard Arabic' may disclose several interesting facts to an interested reader.

Silke Scheible investigates customer's reviews from the interesting point of view: superlatives. Customer's reviews are a very valuable tool representing a feedback, however, what is the relevance of superlatives? How such superlatives contribute really to opinions? Some answers can be found in the article 'The smallest, cheapest, and best: Superlatives in Opinion Mining'.

Not surprisingly, due to the globalisation, today we can find various opinions expressed in tens of different languages. The sentiment classification in Russian and English is presented in the article 'Language-specific Features in Multilingual Sentiment Analysis' by Taras Zagibalov, Katerina Belyatskaya, and John Carroll.

We wish you an interesting reading and the insight into the contemporary research results of the scientific area devoted to applying intelligent machine approach to one of very typically human area: Expressing opinions in natural languages.

On behalf of the Editoral Board,

Jan Žižka
Department of Informatics/SoNet Research Center
Faculty of Business and Economics
Mendel University Brno, Czech Republic

Social Web Mapping Dynamic Analysis of Social Networks

Kristína Machová

Department of Cybernetics and Artificial Intelligence, Faculty of Electrical Engineering and Informatics, Technical University in Košice, Letná 9, 042 00 Košice, Slovakia

Tomáš Rakušinec

Department of Cybernetics and Artificial Intelligence, Faculty of Electrical Engineering and Informatics, Technical University in Košice, Letná 9, 042 00 Košice, Slovakia

Abstract

This paper deals with social networks, their types, and an analysis of their behaviour and properties. The main focus is on dynamic analysis of social networks. This paper is divided into the following parts: social networks (social network types, structural characteristics and data representations) and dynamic analysis of social networks (dynamics of social network of discussion channel, its dynamic visualisation and properties). The paper also describes the results of dynamic analysis of the data, which were collected from the Internet Relay Chat (IRC) social network. These results represent a view of moves in the framework of the social network in its entirety. The IRC network has the character of a discussion channel, which serves for solving user's mainly technical problems from the world of IT and the Unix community.

Key words

social web; web mapping; social networks; social network visualisation; dynamics visualisation; dynamic analysis of social networks

Introduction

The social web can be considered an upgrade of the classic web. The classic web can be illustrated with the idea of a worldwide billboard – anybody can publish some information piece and make it accessible for public inspection on the billboard (anybody who has the necessary skills in web page creation that is – but a considerably greater number of web users have the ability only to access published information). On the other hand, the social web or web 2.0 reinforces social interactions among users and provides an opportunity for the great majority of web users to contribute to web content. It can be said that it increases the number of web content providers. Social interactions among users are enabled by communication within social networks, the possibility to contribute to web discussions and so on. Our work focuses on social networking.

The concept of the 'social network' is frequently used and it represents one important form of the on-line social life of many people. Modern social networks help users to maintain connections and contacts to other users without dependency on a geographical locality. A social network cannot be mapped in its entirety. Mapping of a well-known social network (for example Facebook) is a difficult task when we want to visualise it in a way that guarantees readability. So, our work concentrates on small networks and also on network dynamics visualisation.

There are some known classical methods for social network analysis for offline data processing. The data are collected within a period and data collections obtained usually do not contain a time dimension. Our work is concentrated on dynamic social network analysis. Therefore, dynamic attributes and time ordering of observations connected to the researched social network are a matter of principle. Our aim is to integrate dynamics into social networks analysis and discover some new dynamic relations.

Social networks

A social network is a set of entities (nodes, actors) that are connected. This connection can be represented by one or more types of relationship. According to Han (2003) a social network is a heterogenic and multi-relational data set that can be represented by a graph. Nodes and edges of this graph have attributes and objects (nodes) can be classified into categories. Relations can be oriented and need not be binary only. Social networks can be considered not only in a social, but also in a technological, economic or biological context. Examples of such networks can be an electrical distribution network, telecommunication networks, (computer) virus spreading, collaborative networks (co-authors), citing networks, web networks and so on. According to Repka (2011) the main concepts in social network analysis are: *actor, relation, dyad, triad, subgroup, group and network.*

An *actor* is a social entity in a network. The aim of social network analysis is to make sense of relations between these entities and to analyse the consequences of these relations. An actor can be a discrete individual, an organisation or simply a group of social entities. The actors represent the basic units of the social network in the process of its analysis.

Individual actors are connected to each other by social ties. Such a social tie can be: friendship, respect, preference, association, affiliation, interaction and so on. The collection of social ties between actors (entities) is called a *relation*. The relation $R_v \subseteq \{(A \times A) \times R\}$ between actors of the same type is a binary relation, where A is the set of actors, R is the set of relations and v_x is the intensity of the relation.

A tie between two actors is called a *dyad*. A *triad* arises by extension of the dyad by one more actor. A *subgroup* is an extension of the triad. Subgroups play an important role in the dynamic analysis of social networks. The network dynamics represent the moving of actors within the social network, that is the formation and expiration of subgroups. A group is such a set of actors as belongs to a common and finite set and which are connected with measurable ties. So *network* N can be defined as pair of the set of actors S and the set of all relations R over the set of actors:

$$N = (A, R); A = \{A_1, ..., A_X\}; R = \{R_1, ..., R_Y\}$$

One of the objectives of social networks analysis is identification of cohesive sub-groups and their analysis. Cohesive sub-groups are sub-sets of actors – there are strong, direct, intensive and frequent connections among them. They can be identified on the basis of the number and complexity of mutual connections, on the basis of closeness or readability of other actors of a given group and on the basis of connection frequency between actors.

Social network types

Social networks can be divided into types according various points of view. For example according to the number and type of edges (directed, undirected networks, networks with cycles and networks with multi-edges), according to the means of edge evaluation (weighted, un-weighted, marked and temporal networks), according to the number of actor types (monoecious networks contain only one type of actor, while dioecious networks contain more types of actor) and according to the number of relation types (one-relation networks, multi-relation networks). During real network modelling various network types can be distinguished: small-world phenomenon networks, scale-free networks, social cycle networks, and random and lattice networks.

Real networks are networks, which arise in a process of self-organisation. This self-organisation process can optimise the conservation of network local structure, conservation of good communication between network nodes and resistance to random errors. An example of a real network is the Internet or a network of professional contacts or telecommunication networks.

The great majority of nodes of *small-world phenomenon networks* (Kleinberg, 2000) are not connected to other nodes directly, but they do so from other nodes via a small number of edges. The small-world networks have concise shortest paths, high clustering coefficients and small node separation (these structural characteristics will be described in the next section). An interesting property of these networks is a tendency to contain cliques and n-cliques (because of the high clustering coefficient) and many high degree nodes (see next section). A great number of high degree nodes with can lead to high distribution. The *scale free networks* have distribution of a degree minimally dependent on Power Law. More details can be found in (Barabasi, 2003). *Random and lattice networks* have clustering coefficients close to zero. Its node separation grows slowly with increasing node number (Markošová, 2010). The *social cycle network* model represents networks with feedback (Douglas, 2006). Active nodes (actors) of the feedback network communicate through the network to coordinate the process of joining of new actors and to create new relations. For realisation of this process, three parameters are needed: level of actor activity, distance shortening (how soon the actor fails in the effort to find a suitable partner for new relations) and the range of searching within the network.

Structural characteristics of networks

Some properties of some social networks can be researched with the aid of their structural characteristics. The most important and useful characteristics are: node degree, shortest paths and clustering coefficient (Dorogovtsev, 2002).

The node degree k represents the total number of its connections. In the physics literature, this characteristic is usually called the ‘connectivity’, which has a different meaning in graph theory. The node degree consists of ‘in-degree’ and 'out-degree' $k = k_i + k_o$. The in-degree is the number of incoming edges of a node and the out-degree is the number of outgoing edges. It holds that:

$$P(k) = \sum_{k_i} P(k_i, k - k_i) = \sum_{k_0} P(k - k_0, k_0), \tag{1}$$

where: $P(k)$ is the degree distribution and $P(k_i, k_0$ is the joint in-degree and out-degree distribution. The in-degree distribution $P_i(k_i)$ and the out-degree distribution $P_0(k_0)$ (more obvious are notations $P(k_i)$ and $P(k_0)$ can be stated according the formula (2):

$$P_i(k_i) = \sum_{k_0} P(k_i, k_0), P_0(k_0) = \sum_{k_i} P(k_i, k_0), \tag{2}$$

If a given network has no connections to exterior space, then the average in-degree is equal to the average out-degree, see formula (3):

$$\bar{k_i} = \sum_{k_i, k_0} k_i P(k_i, k_0) = \bar{k_0} = \sum_{k_i, k_0} P(k_i, k_0), \tag{3}$$

The node degree is a local characteristic, but we shall see that degree distribution can determine some important global characteristics of random networks. The average node degree within the whole network represents the network connectivity.

The shortest path is defined as the geodetic distance of two nodes u and v, which represents the shortest one from all possible paths between these two nodes l_{uv}. The shortest path l_{uv} need not be the same as l_{vu}. The average shortest path between all node pairs within the entire network is often called the diameter of a network. It is related to the average separation of pairs of nodes.

This node separation represents another global property of a network – the closeness of nodes. The node separation l is the average shortest distance $d_{min}(a, b)$ of the nodes a and b.

$$l = \frac{1}{N} \sum_{i=1}^{N} d_{min}(a, b) \tag{4}$$

Let us consider a network with undirected edges. Then the number of all possible connections of the nearest neighbours k_i of a node u is ${}^{|k_i}_{2|=\frac{k_i(k_i-1)}{2}}$. Let only E_i are present. Then the clustering coefficient of the node u can be counted according to formula (4).

$$C_i = \frac{2E_i}{k_i(k_i - 1)} \tag{5}$$

The Average clustering coefficient over all nodes of a network is the network clustering coefficient aC.

$$C = \frac{1}{N} \sum_{i=1}^{N} C_i \tag{6}$$

This clustering coefficient represents the probability that the two nearest neighbours of a node are also the nearest neighbours of another node.

Other parameters used in social network analysis are node numbers, edges number, nodes (edges) number in the larger weakly (strongly) connected component in the network, the number of triangles and effective average of the 90th percentile, which is the 90th percentile of the distribution of all shortest path lengths.

Network data and their representation

Social networks are sources of structural, compositional and temporal data. The structural data are obtained from connections between actors (for example friendships between people, business transactions between firms and so on). The structural data can be represented by a quadratic binary matrix (1 represents a connection and 0 represents an absence of connection). The compositional data introduce attributes of network actors, which can be direct (for example interests, age, nationality, country and so on) and indirectly obtained in the process of network analysis by counting (for example node degree, node centrality and so on). These data are static and they can be represented by an orthogonal matrix. Sometimes we are interested in the dynamics of these data, which can be researched in the process of dynamic analysis of social networks. This dynamic analysis uses temporal data, which describe the dynamic properties of a social network. The temporal data represent a time change and its time evidence.

Dynamic analysis of social networks

The dynamic analysis focuses on time changes in a researched social network. Thus the time dimension must be added into static data. Within dynamic analysis, we monitor the network development in some direction. The dynamic analysis uses visualisation methods, structural characteristics determination, machine learning methods, multi-agents modelling and combinations of these. The main problems of dynamic analysis according to Repka (2011) are:

1. searching for a suitable metric for dynamic analysis;
2. prediction of changes in social networks;
3. developing of algorithms for social network monitoring;
4. dynamic visualisation of social networks;
5. research of social networks and their characteristics in time.

Our work focuses on the last two problems. The static as well as dynamic analysis searches for suitable metrics for network description. These metrics can be used to solve the actual problem of searching for and identifying authoritative actors in a social network or in some group of the network.

In general, dynamic changes represent the creation and destruction of nodes. In real networks, the destruction of a node is often inappreciable. Such a network is called a growing network (Barabasi, 1999). Network growth can be realised both by random addition of actors and by preferential addition of actors.

The random addition of an actor is a process starting from one actor. Another actor is added to an existing actor by creating one connection to it in each time unit. Thus, the actors can be marked by the time of their coming into a network. At time t, the network has exactly t actors. So the average node degree of the node, which came into the network at time s and which we monitor at time $t+1$ is:

$$k(s, t+1) = k(s,t) + \frac{1}{t+1} \tag{7}$$

This equation can be solved and the result of its solution is formula (8):

$$k(s,t) = 1 - ln(\frac{s}{t}) \tag{8}$$

The preferential addition of actors similarly starts from one actor. In each time unit, a new actor is connected to the existing actors by creating one or more connections to existing actors in the network. The new actor selects the existing actor for connection that has the highest node degree (the highest number of existing connections). Each existing connection contributes to a node degree increase for two actors. The average node degree of the node, which came into the network at time s and which we monitor it at time t is:

$$k(s,t) = (\frac{t}{s}^{-2}). \tag{9}$$

The distribution function of node degrees is:

$$P(k, t \to \infty) = 2k^{-\gamma}, \gamma \ = 3. \tag{10}$$

The network with this distribution function has actors with a low number of connections as well as actors with a high number of connections to other actors. Such a network is called a scale-free network (Barabasi, 2003) and the model described is the Barabasi-Albert model.

Dynamics of the social network of a discussion channel

During the dynamic analysis of a network, our attention was focused on the social network IRC (Internet Relay Chat), which was created on the basis of interactions between users of a discussion channel located on the web. The advantage of this selection is that users of this discussion channel continuously communicate between each other long time, because of themes of their communication is always actual for them. These themes are related to technical problem solving, to technical news discussion or to common communication. The IRC social network is created continuously in the process of communication between users. So it is a naturally dynamic system suitable for dynamic analysis. The dynamics of this system is created by emergence. The dynamics of this discussion channel are not created purposelessly by the user but are related to the actual load of this channel according to the real need to solve some technical problems that have arisen.

From the beginning our tests concentrated on the percentage load of the researched discussion channel during particular phases. Each phase is represented by one hour of the day. Figure 1 illustrates the percentage average loading of the discussion channel in particular hours of the day.

Figure 1: Average percentage of IRC discussion channel loading. Each column represents one hour of the day. Data were collected over 211 days.

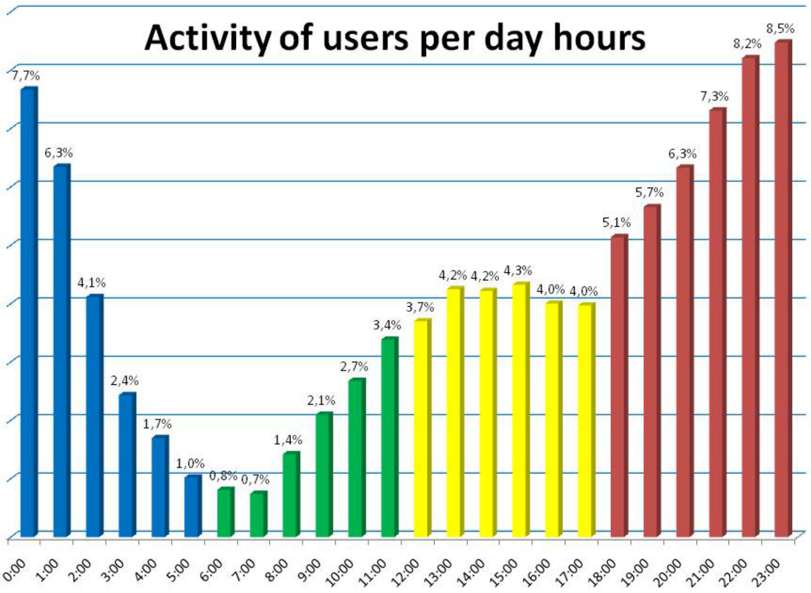


Table 1: Number of users' communications in a certain time period – each hour of the day.

Day hour Number	00:00-00:59 128,242	01:00-01:59 106,138	02:00-02:59 68,807	03:00-03:59 40,686	04:00-04:59 28,396	05:00-05:59 17,086
Day hour Number	06:00-06:59 13,563	07:00-07:59 12,391	08:00-08:59 23,755	09:00-09:59 35,083	10:00-10:59 44,776	11:00-11:59 56,598
Day hour Number	12:00-12:59 61,863	13:00-13:59 71,047	14:00-14:59 70,561	15:00-15:59 72,301	16:00-16:59 66,927	17:00-17:59 66,373
Day hour Number	18:00-18:59 85,975	19:00-19:59 94,583	20:00-20:59 105,862	21:00-21:59 122,248	22:00-22:59 137,262	23:00-23:59 141,740

Figure 2: The first iteration of the visualisation process (initialising set of social network diagram with two entities from October 09, 2009).

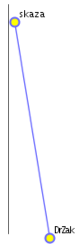
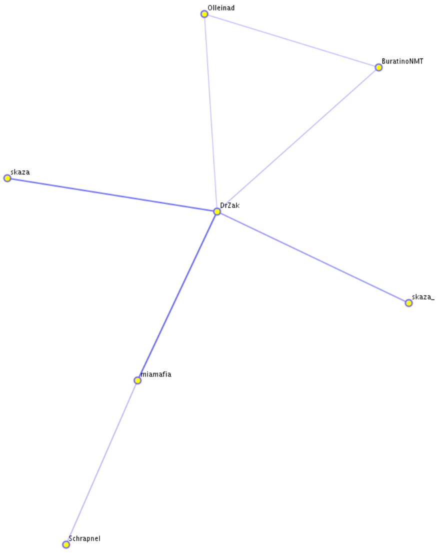


Figure 3: The 42nd iteration of the visualisation process (the social network diagram is growing – October 09, 2009).



The Figure 1 shows that the highest activity on the IRC discussion channel was at night in the evening. Table 1 gives the exact number of iterations spotted. The iterations represent communication activities between users per certain day hour. These data were collected over 211 days.

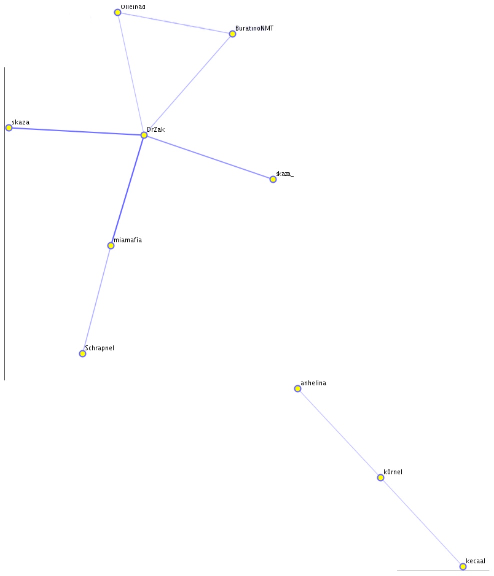
The total number of users' communications over all 211 days was 1 672 263 iterations. Activities on the discussion channel were observed with the aid of the autonomous robot Eggdrop (Eggdrop, 2011). The robot fulfils the role of a manager of users. It provides some added functionalities for operators, moderators and functionalities for safety measures.

Dynamic visualisation of social network

We have tried to visualise the IRC social network in graphical way. The graphical visualisation of the social network is a view of an actual running discussion. Each completed discussion is removed from the graph and each new discussion is added to graph by adding new nodes and edges. Running discussions are depicted by a darkening flow – line, as can be seen in the video 'czechoslovakia.wmv' available on 'http://hron.fei.tuke.sk/ rakusinec/ahReic3E/'.

This video represents running discussions and their cardinality changing over time. Such a view of a social network has a high computing complexity and with time demands on disc space increase. One possible solution to this problem is the creation of an animation that reflects only changes in time. The graph visualisation is modified only at discrete times when a change was noticed. In this way, we transformed the previously mentioned animation 'czechoslovakia.wmv' into a series of

Figure 4: The 44th iteration of the visualisation process (the social network diagram has split into two isolated parts – October 09, 2009).



8497 graph visualisations. The figures from Fig. 2 to Fig. 6 were selected from this series.

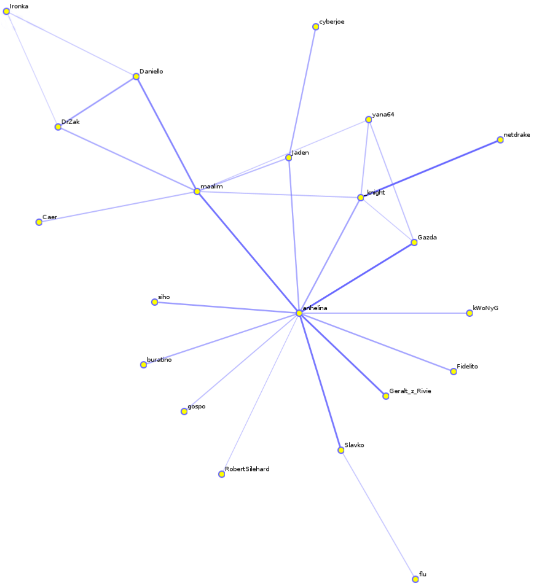
The dynamics of a social network or its evolution can be divided into two phases: an initialising phase and a monitoring phase.

The *initialising phase* starts building the social network diagram. The first interaction between the two first actors is denoted into diagram, as can be seen in Figure 2. At the beginning of building the social network diagram, the greatest changes in social network structure are denoted.

The *phase of social network monitoring* is characterised by oscillation around the stabilised graph. The oscillations depend on setting the value of a parameter that controls the time, and non-active nodes are removed from the graph after their expiration. Too low a value of this parameter enables social network monitoring nearly on-line, because the process of network development is accelerated. In this case a node is removed strictly, after very short period of inactivity. On the other hand, too high a value of this parameter causes a node to be removed after a very long period of inactivity. An extremely high value of the parameter can mean that no node is removed from graph and such a graphical visualisation can provide us with a global view of all interactions between users from the first interaction to the actual, or last one. However such a complex visualisation gradually loses its information value and readability, as illustrated in Figure 7.

The graphical visualisation of a social network and also its dynamic analysis based on graphical visualisation can be overly complex and complicated. This was our reason for focusing on the analysis of social network properties.

Figure 5: The 1778th iteration of the visualisation process (the social network diagram has become more complicated – Jun 09, 2010).



Dynamic properties of social networks

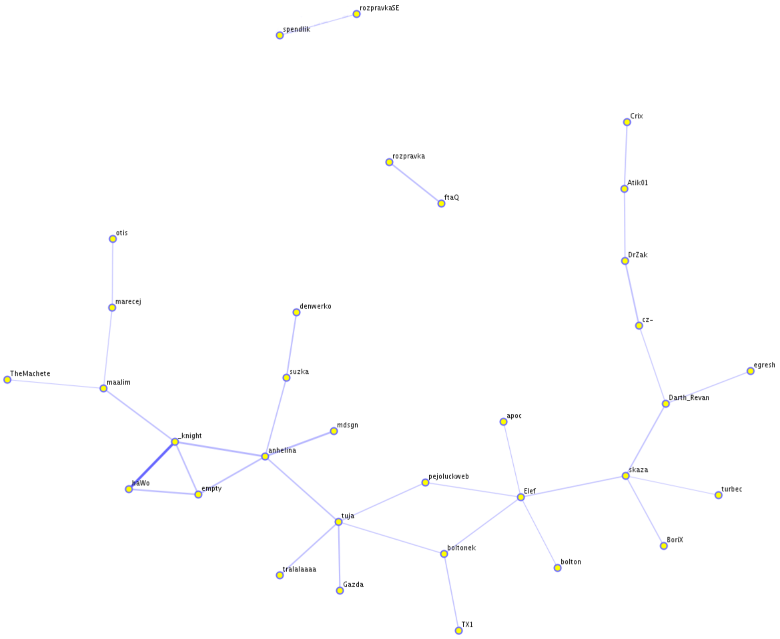
Some structural characteristics of networks were introduced within section 2.2. These static properties can be studied in their dynamics. We concentrated our attention on one of them – the node degree *k*, which is the most basic parameter of the investigated point (node, entity, user or actor) from the point of view of dynamic visualisation and analysis of social networks. The node degree *k* represents the number of all the nearest neighbours of the given node (see section 2.2).

Our experiments are related to a closed social network – the IRC discussion channel. All nodes communicate only to other nodes in the same network. The average in-degree of the node is equal to the average out-degree (see formula (3)), because the investigated network is a discussion channel, which works on the basis 'question – answer'. The set of users that did not get an answer is small and can be ignored.

We tried to research the average user-discussant and came to an average distribution of the node (user) degree depending on time. This node degree distribution is illustrated in Figure 8.

It can be seen in Figure 8 that the average node degree oscillates between values of 2 and 8. It is a relatively common value, which represents the communicative abilities of one user (actor, discussant). This value is depends also on the discussion channel load and the average number of active users-discussants during the day.

Figure 6: The 8497th iteration of the visualisation process (last social network diagram – December 05, 2010).



Conclusion

The purpose of this work was research into the dynamics of the IRC social network. We have analysed data, which were collected by PieSpy (PieSpy, 2011). The total number of collected screens was 181310. It takes up 5 GB of disc space. The social network was monitored over enough sufficient time (from October 9, 2009 to December 05, 2010) to trace permutations in the social network organisation. Clusters creation, central nodes creation and small isolated islands of small numbers of nodes were denoted.

Dynamic processes originate in this social network through emergence. This means that network users do not communicate with other users for the purpose of social network building or to gain connections, but to solve technical problems. This social network is the result of independent communication on a public level. As in any common society so in this social network the central nodes are more interesting. The central nodes – hubs – are typical, with a high number of connections to other nodes-entities. These central nodes represent very active actors, who spend more time on the discussion channel, or authoritative users, who are the best orientated in problem domains. The identification of such authoritative actors is interesting problem for us for future research.

Another promising research field is discussion analysis (Lukáč, 2008). This problem is related to discussion channels as well, but it concentrates on opinion analysis rather than dynamics analysis. This opinion analysis can be semantically

Figure 7: Overly massive visualisation of a social network (passed on from Ryze Business Networking)

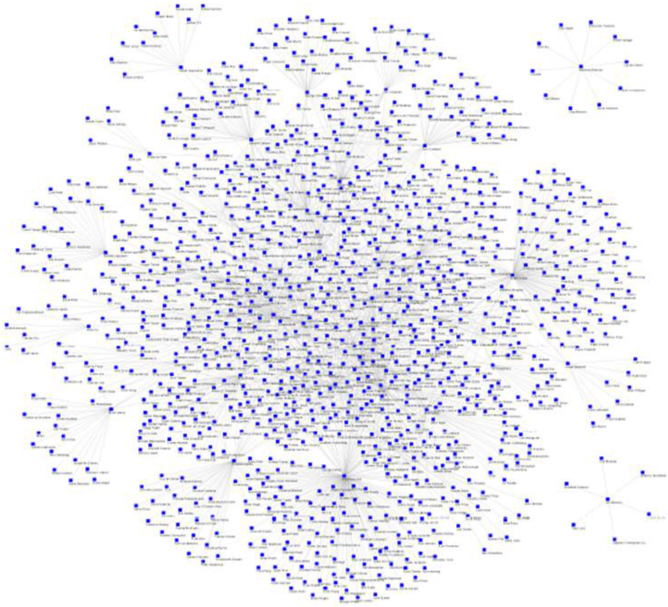
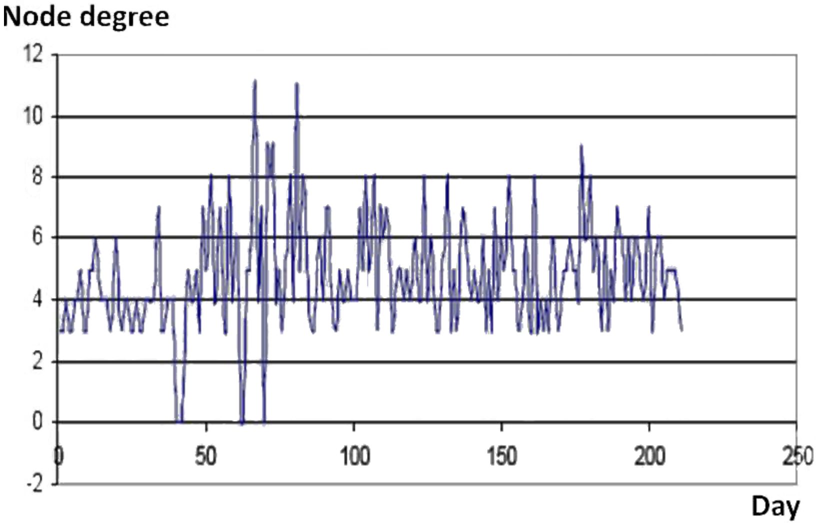


Figure 8: The distribution of the average node degree of the social network depending on time.



enhanced, which can create hybrid access to semantic and social web re-approach. This access represents social web mining from content. Social web mining from a structure can also be interesting, as introduced in Lukáč (2010) with a focus on a social network of authors and the tracking of influential concepts spreading down this network.

References

Barabasi, A. L. (2003): Scale-free Network. *Scientific American*, 288, pp. 60–69.

Barabasi, A. L., Albert, R. (1999): Emergence of scaling in random networks. *Science*, 286, pp. 509–512.

Dorogovtsev, S. N., Mendes, J. F. F. (2002): Evolution of Networks. *Adv. Phys*, Vol. 51, pp. 1079–1187.

Douglas, R. a kol. (2006): Generative Model for Feedback Networks. In: *Physical Review E.* , pp. 016119-1–016119-8.

EGGDROP. (2011): The world's most popular open source Internet Relay chat bot [on-line]. [cit. 2011-04-11]. Available at: http://www.omgirc.com/index.php?itemid=26.

Hann, W. P. (2003): Hyperlink Network Analysis: A New Method for the Study of Social Structure on the Web Royal Netherlands Academy of Arts & Sciences, Amsterdam, The Netherlands. *Connections*, Vol. 25, No. 1, pp. 49–61.

Kleinberg, J. (2000): The Small-world Phenomenon: An Algorithmic Perspective. In: *STOC '00: Proceedings of the thirty-second annual ACM symposium on Theory of computing.* New York : The Association for Computing Machinery, pp. 163–170. (ISBN 1-58113-184-4.)

Lukáč, G. (2010): A Proposal for an Approach to Extracting Conceptual Descriptions of Hyper-linked Text Documents. *Int. J. on Social Media MMM: Monitoring, Measurement and Mining*, Vol. 1, No. 1, pp. 98 to 101. (ISSN 1804-5251.)

Lukáč, G., Butka, P., Mach, M. (2008): Semantically-enhanced extension of the discussion analysis algorithm in SAKE. In: *Proceedings of the SAMI 2008.* Budapest : Budapest Tech, pp. 241–246.

Markošová, M. (2010): Networks Dynamics. In: Kvasnička a kol. (ed.) *Artificial Intelligence and Cognitive Sciences II.* Bratislava (SK) : STU, pp. 321–377.

PieSpy. (2011): PieSpy Social Network Bot. [on-line]. [cit. 2011-04-14]. Available at: http://www.jibble.org/piespy/.

Repka, M. (2011): *Analysis of selected types of social networks.* Košice (SK) : Technical University in Košice, 75 pp.

RBN. (2011): Ryze Business Networking [on-line]. [cit. 2011-04-14]. Available at: http://www.ryze.com/index.php.

This work was supported by the Slovak Grant Agency of the Ministry of Education and Academy of Sciences of the Slovak Republic within project 1/0042/10 'Methods for identification, annotation, search, access and composition of services using semantic metadata in support of selected process types' (50%). This work is also the result of the implementation of the project Development of the Centre of Information and Communication Technologies for Knowledge Systems (project number: 26220120030) supported by the Research & Development Operational Program funded by the ERDF (50%).

Linguistically-Motivated Subjectivity and Sentiment Annotation and Tagging of Modern Standard Arabic

Muhammad Abdul-Mageed
Department of Linguistics , and School of Library Information Science, Indiana University, Bloomington, 1320, E. Tenth Street Bloomington, IN, tel. U.S.A.+001-8128552018
e-mail: `mabdulma@indiana.edu`

Mona Diab
Center for Computational Learning Systems, Columbia University, NY, 475, Riverside Drive, Suite 850 Interchurch Center, MC 7177, tel. U.S.A.+001-2128701290
e-mail: `mdiab@ccls.columbia.edu`

Abstract

There has been recently a swelling interest in the area of *Subjectivity and Sentiment Analysis (SSA)*. However, only few attempts have been made to build SSA systems for *morphologically-rich languages (MRL)*. In addition, although there is a number of studies reporting annotation of various data sets for SSA, most of these studies are not strongly motivated by existing linguistic theory. In the current paper, we aim at partially bridging these two gaps in the literature. More specifically, we (1) provide a detailed description of a highly successful SSA system built for Modern Standard Arabic (MSA), a MRL, and (2) provide linguistically-motivated annotation guidelines for SSA.

Key words

subjectivity; sentiment, sentiment analysis; social meaning; Web mining; Arabic; morphologically-rich languages

Introduction

Subjectivity in natural language refers to aspects of language used to express opinions, feelings, evaluations, and speculations (Banfield, 1982; Wiebe, 1994) and, as such, it incorporates sentiment. The process of subjectivity classification refers to the task of classifying texts as either objective (e.g., *The NATO bombed Gaddafi troops*) or *subjective*. Subjective text is further classified with *sentiment* or *polarity*. For sentiment classification, the task refers to identifying whether a subjective text is *positive* (e.g., *The Egyptian and Tunisian revolutions are impressive!*), *negative* (e.g., The bloodbaths in Syria are horrifying!), *neutral* (e.g., *Saleh of Yemen may step down soon.*), or, sometimes, *mixed* (e.g., *I adore this camera, but it is prohibitively expensive*).

Two main issues arise in SSA. First, available approaches to labeling data for SSA vary considerably. Although our specific goal here is not to describe the various approaches, nor to provide a single standardized approach, we do seek to show how annotation studies within SSA can be inspired by existing linguistic theory. More specifically, by describing our efforts to label a specific corpus for SSA and summarizing our linguistically-motivated guidelines for the task, we hope to trigger a stronger tie between existing linguistic theory and efforts to label data for social meaning tasks such as SSA.

Second, in spite of the flurry of research within the area of SSA, only few attempts have been made to build SSA systems for *morphologically-rich languages (MRL)* (i.e., languages in which significant information concerning syntactic units and relations are expressed at the word-level (Tsarfaty et al., 2010). We thus also aim at partially bridging this gap in the literature by reporting some of our recent work on Arabic, a very morphologically-complex language. We present work that investigates the role of morphology in SSA systems. We investigate Modern Standard Arabic (MSA), a morphologically-rich variety of Arabic, e.g., (Diab, 2007; Habash, Rambow and Roth, 2009). More specifically, we explore the task of sentence-level SSA on (MSA) texts from the news genre. We run experiments on three different pre-processing settings based on tokenized text from the Penn Arabic Treebank (PATB) (Maamouri et al., 2004) and employ both language-independent and Arabic-specific, morphology based features. Our work shows that explicitly using morphology-based features in our models improves the system's performance. We provide a detailed analysis of the performance of the system. We also measure the impact of using a wide coverage polarity lexicon and show that using a tailored resource results in significant improvement in classification performance.

The rest of the paper is organized as follows: In Section , we introduce the news genre. In Section , we overview the annotation process and the categories of each annotation task. In Section , we describe the (linguistics) insights driving our annotation. In Section , we provide examples from each SSA category. In Section , we describe our approach, including polarity lexicon, and automatic classification methodology. In Section , we present the results and evaluation. In Section , we provide an error analysis. Section is about related work and Section is the conclusion.

Subjectivity and Sentiment in the News

The bulk of SSA work has been performed on highly subjective, user-generated data such as blogs and product or movie reviews. In these data, authors tend to express their opinions quite explicitly Abdul-Mageed and Diab, 2011; Balahur and Steinberger, 2009). Consequently, data belonging to genres traditionally viewed as less subjective, like the news genre, have not receieved much attention. News, however, plays an instrumental role in modern societies (e.g., as an influencer of the social construction of reality Fowler, 1991; Chouliaraki and Fairclough, 1999; Wodak and Meyer, 2009).

In addition, news-making is increasingly becoming an interactive process in which lay Web users are getting more and more involved (Abdul-Mageed, 2008): News-makers reproduce some of the views of their readers (e.g., by quoting them) and they devote full stories to the interactions of web users on social media outlets. Although subjectivity in news articles has traditionally tended to be implicit, this growing trend to foster interactivity and more heavily report communication of internet users within the body of news articles is likely to make expression of subjectivity in news articles more explicit. Furthermore, the fact that news stories

	OBJ	**S-POS**	**S-NEG**	**S-NEUT**	**Total**
OBJ	1192	21	57	11	1281
S-POS	47	439	2	3	491
S-NEG	69	0	614	6	689
S-NEUT	115	2	9	268	394
Total	1423	462	682	288	2855

Table 1: Agreement for SSA sentences

have their own biases (e.g., hiding agents behind negative or positive events via use of passive voice, variation in lexical choice) has been pointed out (e.g., van Dijk, 1988) and hence the news genre is definitely not as objective as it may seem to some.

Data Annotation

Two graduate-level native speakers of Arabic annotated 2855 sentences from Part 1 V 3.0 of the Penn Arabic TreeBank (PATB) (Maamouri et al., 2004). The sentences make up the first 400 documents of that part of PATB amounting to a total of 54.5% of the PATB Part 1 data set. The task was to annotate MSA news articles at the sentence level. Each article has been processed such that coders are provided individual sentences to label. We prepared annotation guidelines for this SSA task focusing specifically on the newswire genre. We summarize the guidelines next, illustrating related and relevant literature.

Subjectivity and Sentiment Categories

For each sentence, each annotator assigned one of 4 possible labels: (1) Objective (OBJ), (2) Subjective-Positive (S-POS), (3) Subjective-Negative (S-NEG), and (4) Subjective-Neutral (S-NEUT). We followed Wiebe, Bruce and O'Hara (1999) in operationalizing the subjective vs. the objective categories. In other words, if the primary goal of a sentence is perceived to be the objective reporting of information, it was labeled OBJ. Otherwise, a sentence would be a candidate for one of the three subjective classes.[1] Table shows the contingency table for the two annotators' judgments. Overall agreement is 88.06%, with a Kappa (*k*) value of 0.38.

By way of illustration, a sentence such as "The Prime Minister announced that he will visit the city, saying that he will be glad to see the injured", has two authors (the story writer and the Prime Minister indirectly quoted). Accordingly to our guidelines, this sentence should be annotated S-POS tag since the part related to the person quoted (the Prime Minister) expresses a positive subjective sentiment,

[1] It is worth noting that even though some SSA researchers include subjective mixed categories, we only saw such categories attested in a negligible percent of the sentences which is expected since our granularity level is the sentence. If we are to consider larger units of annotation, we believe mixed categories will become more frequent. Thus we decided to tag the very few subjective mixed sentences as S-NEUT.

Domain	**# of Cases**
Politics	1186
Sports	530
Military & political violence	435
Disaster	228
Economy	208
Culture	78
Light news	72
Crime	62
This day in history	56
Total	**2855**

Table 2: Domains

"glad" which is a *private state* (i.e., a state that is not subject to direct verification) (Quirk et al., 1974).

Domain Annotation and Categories

The same two annotators also manually assigned each sentence a domain label. The domain labels are from the news genre and are adopted from (Abdul-Mageed, 2008). The set of domain labels is as follows: {*Light news, Military and political violence, Sport, Politics, Crime, Economy, Disaster, Arts and culture, This day in history*}. Table illustrates the number of sentences deemed for each domain. Domain annotation is an easier task than subjectivity annotation. Inter-annotator agreement for domain label assignment is at 97%. The two coders discussed differences and a total agreement was eventually reached. Coders disagreed most on cases belonging to the *Military and political violence* and *Politics* domains. For example, the following is a case where the two raters disagreed (and which was eventutally assigned a *Military and political violence* domain):

طلب رئيس الوزراء السابق في جزر فيدجي ماهندرا شودري الذي أطيح به

في ١٩ أيار مايو إثر حركة انقلابية، اليوم السبت باعادة حكومته إلى السلطة.

Transliteration: Tlb r}ys AlwzrA' AlsAbq fy jzr fydjy mAhndrA $wdry Al*y OTyH bh fy 19 OyAr mAyw Ivr Hrkp AnqlAbyp, Alywm Alsbt bIEAdp Hkwmth IlY AlslTp.
English: Former Prime Minister of Fiji Mahendra Chaudhry, who was ousted in May 19 after a revolutionary movement, asked on Saturday to return to office.

(Linguistic) Insights Driving Annotation

Good & Bad News vs. S-POS & S-NEG

In general, news can be either good or bad. For instance, whereas "Five persons were killed in a car accident" is bad news, "It is sunny and warm today in Chicago" is good news. Our coders were instructed not to consider *good* news *S-POS* nor

bad news *S-NEG* if they think the sentences expressing them are objectively reporting information. Thus, bad news and good news can be OBJ as is the case in both examples. Indeed, this specific nuance makes news-focused SSA a difficult task.

Role of Epistemic Modality

Epistemic modality serves to reveal how confident writers are about the truth of the ideational material they convey (Palmer, 1986). Epistemic modality is classified into *hedges* and *boosters*. *Hedges* are devices like *perhaps* and *I guess* that speakers employ to reduce the degree of liability or responsibility they might face in expressing the ideational material. *Boosters*[2] are elements like *definitely, I assure that*, and *of course* that writers or speakers use to emphasize what they really believe. Both hedges and boosters can (1) turn a given unit of analysis from objective into subjective and (2) modify polarity (i.e., either strengthen or weaken it). Consider, for example, the sentences (1) "Gaddafi has murdered hundreds of people", (2) "Gaddafi may have murdered hundreds of people", and (3) "Unfortunately, Gaddafi has definitely murdered hundreds of people". While (1) is OBJ, since it lacks any subjectivity cues, (2) is S-NEUT because the proposition is not presented as a fact but rather is softened and hence offered as subject to counter-argument, (3) is a strong S-NEG (i.e., it is S-NEG as a result of the use of "unfortunately", and *strong* due to the use of the booster *definitely*). Our annotators were explicitly alerted to the ways epistemic modality markers interact with subjectivity.

Role of Perspective

Sentences can be written from different *perspectives* (Lin et al., 2006) or points of view. Consider the two sentences (1) "Israeli soldiers, our heroes, are keen on protecting settlers" and (2) "Palestinian freedom fighters are willing to attack these Israeli targets". While sentence (1) is written from an Israeli perspective, sentence (2) is written from a Palestinian perspective. The perspective from which a sentence is written interplays with how sentiment is assigned. Sentence (1) can usually be considered positive from an Israeli perspective, yet the act of protecting settlers is, more often than not, viewed as negative from a Palestinian perspective. Similarly, attacking Israeli targets may be positive from a Palestinian vantage point, but will perhaps be negative from an Israeli perspective. Coders were instructed to assign a tag based on their understanding of the type of sentiment, if any, the author of a sentence is trying to communicate. Thus, we have tagged the sentences from the perspective of their authors. As it is easy for a human to identify the perspective of an author (Lin et al., 2006), this measure facilitated the annotation task. Thus,

[2] Polanyi and Zaenen(2006) call these *intensifiers*.

knowing that the sentence (1) is written from an Israeli perspective, the annotator assigns it a S-POS tag.

Role of Illocutionary Speech Acts

Occurrences of language expressing (e.g. *apologies, congratulations, praise*, etc. are referred to as *illocutionary speech acts* (ISA) (Searle, 1976). We strongly believe that ISAs are relevant to the expression of sentiment in natural language. For example, the two categories *expressives* (e.g., congratulating, thanking, apologizing) and *commiss ives* (e.g., promising) of Searle's (1976) taxonomy of ISAs are specially relevant to SSA. In addition, Bach and Harnish (1979) define an ISA as a medium of communicating attitude and discuss ISAs like *banning, bidding, indicting, penalizing, assessing* and *convicting*. For example, the sentence "The army should never do that again" is a *banning* act and hence is S-NEG. Although our coders were not required to assign ISA tags to the sentences, we have brought the the concept of ISAs to their attention as we believe a good understanding of the concept does facilitate annotating data for SSA.

Role of Annotator's Background Knowledge

The type of sentiment expressed may vary based on the background knowledge of an annotator/reader (Balahur and Steinberger, 2009). For example, the sentence "Secularists will be defeated", may be positive to a reader who opposes secularism. However, if the primary intention of the author is judged to be communicating negative sentiment, annotators are supposed to assign a S-NEG tag. In general, annotators have been advised to avoid interpreting the subjectivity of text based on their own economic, social, religious, cultural, etc. background knowledge.

Examples of SSA categories from MSA news

We illustrate examples of each category in our annotation scheme. We also show and discuss examples for each category where the annotators differed in their annotations. Importantly, the two annotators discussed and adjudicated their differences.

Objective Sentences

Sentences where no opinion, sentiment, speculation, etc. is expressed are tagged as OBJ. Typically such sentences relay factual information, potentially expressed by an official source, like examples 1-3 below:

(١) ويبلغ عدد المشردين فى كنتية لوس انجلس نحو ٨٤ الف شخص.

Transliteration:[3] wyblg Edd Alm$rdyn fy kwntyp lws Onjlys nHw 84 Olf $xS.
English: The number of homeless in Los Angeles County is about 48 thousand.

[3] We use here Buckwalter transliteration www.qamus.org.

(٢) طهران ١٥ -٧ (أف ب) -وقع ١٦ انفجارا مساء اليوم السبت في وزارة
الاستخبارات حيث استدعيت العديد من سيارات الاسعاف كما أكد شاهد عيا
لوكالة فرانس برس.

Transliteration: ThrAn 15-7 (A f b) - wqE 16 AnfjArA msA' Alywm Alsbt fy wzArp AlAstxbArAt Hyv. AstdEyt AlEdyd mn syArAt AlIsEAf kmA Okd $Ahd EyAn lwkAlp frAns brs.
English: Tehran 15-7 (AFP) - An eye witness affirmed to AFP that 16 explosions occurred late Saturday at the Ministry of Intelligence where many ambulances were summoned.

(٣) أعلن السائق الأيرلندي أيدي أيرفاين (جاغوار) انسحابه من سباق جائزة
النمسا الكبرى.

Transliteration: AEln AlsA}q AlIyrlndy Iydy IyrfAyn (jAgwAr) {nsHAbh mn sbAq jA}zp AlnmsA AlkbrY.
English: The Irish driver Eddie Irvine (Jaguar) announced his withdrawal from the Austrian Grand Prix.

Examples 1-3 show that objective sentences can have some implicitly negative words/phrases like انتعشت الآمال“{nsHAb” (“withdrawl”). In addition, although these 3 examples convey *bad* news, they are annotated with an OBJ tag since the sentences are judged as facts, although one annotator did initially tag example 1 as S-NEG before it was resolved later.

Subjective Positive Sentences

Sentences that were assigned a S-POS tag included ones with positive *private states* (Quirk et al., 1974; i.e., states that are not subject to verification). Examples 4 and 5 below are cases in point where the phrase انسحاب“AntE$t Al|mAl” (“hopes revived”) and the word

طمئنان'

“TmnAn” (“relief”) stand for unverifiable private states:

(٤) وانتعشت الامال بالافراج عن الرهائن في الساعات ال --٢٤ الاخيرة مع
تدخل ليبيا.

Transliteration: wAntE$t Al*vert*mAl bAlIfrAj En AlrhA}n fy AlsAEAt Al 24 AlAxyrp mE tdxl lybyA.
English: Hopes for the release of hostages revived in the last 24 hours with the intervention of Libya.

(٥) و أبدى صلات حسن اطمئنانه إلى عودة النظام والاستقرار إلى بلاده.

Transliteration: wAbdY SlAt Hsn TmnAnh IlY Ewdp AlnZAm wAlstqrAr IlY blAdh.

English: Silaat Hasan expressed relief for the return of order and stability to his country.

The subtle nature of subjectivity as expressed in the news genre is reflected in some of the positive examples, especially in directly or indirectly quoted content when quoted people express their emotion or support their cause (via e.g., using modifiers). For instance, the use of the phrases من اجل نهضة الصومال“mn Ajl nhDp AlSwmAl” (“for the advancement of Somalia”) in example 6 below turn what would have otherwise been OBJ sentences into S-POS sentences. Again, one annotator initially tagged example 8 as OBJ):

(٦) دعا الرئيس الصومالي مساء أمس السبت الدول المانحة وخصوصا أعضاء
الجامعة العربية والاتحاد الأوروبي إلى تقديم مساعدات إلى بلاده \"من أجل
نهضة الصومال \".

Transliteration: dEA Alr}ys AlSwmAly msA’ Ams Alsbt Aldwl AlmAnHp wxSwSA AEDA’ AljAmEp AlErbyp wAl{tHAd AlAwrwby IlY tqdym msAEdAt IlY blAdh “mn Ajl nhDp AlSwmAl”.
English: The Somali President, on Saturday evening, called on the donor countries, especially members of the Arab League and the European Union, to provide assistance to his country “for the advancement of Somalia”.

Quoted content sometimes was in the form of *speech acts* (Searle, 1975). For example, (7) is an *expressive speech act* where the quoted person is thanking another party:

(٧) [وأضاف:] \"شكرا من أعماق قلبي لهذا الشرف الذي يمتد مدى الحياة. \"

Transliteration: [wADAf:] “”$krA mn AEmAq qlby lh‘*A Al$rf Al*y ymtd mdY AlHyAp”.
English: [He added:] Thank you from all my heart for this life-long honor.

Subjective Negative Sentences

Again, the more explicit negative content was found to be frequent in sentences with quoted content (as is illustrated in examples 8 and 9). (8) shows how the S-NEG S-POS sentiment can be very strong as is illustrated by the use of the noun phrase إصرار شيطاني “ISrAr $yTAny” (“diabolical insistence”):

(٨) ورد أحد محامي أندريوتي جيواكينو على قرار النيابة في باليرمو واصفا إياه
بأنه \"إصرار شيطاني \"من قبل الاتهام.

Transliteration: wrd AHd mHAmy Andrywty jywAkynw sbAky ElY qrAr AlnyAbp fy bAlyrmw wASfA IyAh bAnh “ISrAr $yTAny” mn qbl AlAthAm.

English: One of lawyers of Andreotti Jjoaquino responded to the prosecutor's decision in Palermo, describing it as a "diabolical insistence" on the acusser's part.

Speech acts have also been used to express negative sentiment. For example, (14) is a direct quotation where a political figure denounces the acts of hearers. The speech act is intensified through the use of the adverb

(٩) وقال شارون من منصة الكنيست متوجها إلى نواب حزب العمل: "\القد
تخليتم حتى عن القسم الأكبر من المدينة القديمة "\.

"HtY" ("even"): حتى

Buckwalter: wqAl $Arwn mn mnSp Alknyst mtwjhA AlY nwAb Hzb AlEml "lqd txlytm HtY En Alqsm AlAkbr mn Almdynp Alqdymp."
English: Sharon, addressing Labour MPs from the Knesset, said: "You have even abandoned the biggest part of the old city".

Majority of the sentences pertaining to the *military and political violence* domain were OBJ, however, some of the sentences belonging to this specific domain were annotated S-NEG. News reporting is supposed to be objective, story authors sometimes used very negative modifiers, sometimes metaphorically as is indicated in (10). Example 10, however, was labeled OBJ by one of the annotators, and later agrrement was reached that it is more of an S-NEG case.

(١٠) وكان شهر تموز (يوليو) دمويا بشكل خاص مع سقوط نحو ٣٠٠ قتيل.

Transliteration: wkAn $hr tmwz ywlyw dmwyA b$kl xAS mE sqwT nHw 300 qtyl.
English: The month of July was especially bloody, with the killing of 300 people.

Subjective Neutral Sentences

Some of the S-NEUT cases were speculations about the future, as is illustrated by sentences 11 and 12:

(١١) ويتوقع أن يعود إلى الولايات المتحدة في ٢٥ تموز (يوليو).

Transliteration: wytwqE An yEwd IlY AlwlAyAt AlmtHdp fy 25 tmwz (ywlyw).
English: And he is expected to return to the United States on July 25.

(١٢) وكل الموءشرات تفيد أن هذا الوضع لن يتغير بعد الانتخابات.

Transliteration: wkl AlmW$rAt tfyd In h'*A AlwDE ln ytgyr bEd AlAntxAbAt.
English: All indications are that this situation will not change after the elections.

Hedges were also used to show cautious commitment to propositions, and hence turn OBJ sentences to S-NEUT ones. Sentences (13) and (14) are examples, with the occurrence of the hedge trigger word يبدو "ybdw" ("it seems") in (13) and على الأرجح "ElY AlArjH" ("it is most likely") in (14):

Word	POS	Surface form	Lemma	Stem	Gloss
AlwlAyAt	Noun	A+lwlAyAt	Al+wlAyp	Al+wlAy	the states
ltblgh	Verb	l+tblg+h	l+Oblg+h	l+blg+h	to inform him

Table 3: Examples of word lemmatization settings

(١٣) و يبدو أن التكتم الذي أحاط بزيارة بيريز إلى أندونيسيا كان يهدف إلى
تفادي إثارة ردود فعل معادية في البلاد.

Transliteration: w ybdw An Altktm Al*y AHAT bzyArp byryz AlY AndwnysyA kAn yhdf AlY tfAdy AvArp rdwd fEl mEAdyp fy AlblAd.
English: It seems that the secrecy surrounding Peres's visit to Indonesia was aimed at avoiding negative reactions in the country.

(١٤) وعلى الأرجح أن قبطان الغواصة أعطى الأمر ب -إطفاء كل الآلات على
متنها.

Transliteration: wElY AlArjH An qbTAn AlgwASp AETY AlAmr bATfA' kl AlAlAt ElY mtnhA.
English: Most likely the submarine's captain ordered turning off all the machines on board.

Approach

Polarity Lexicon

To the best of our knowledge, there is no publicly available polarity lexicon for Arabic. Accordingly, we manually created a lexicon of 3982 adjectives extracted from the first four parts of the PATB. Each adjective was labeled with one of the tags in the set {*positive, negative, neutral*}.

Automatic Classification Methodology

Settings: We run experiments on gold-tokenized text from PATB. We adopt the PATB+Al tokenization scheme, where proclitics and enclitics as well as the definite article "Al" are segmented out from the stem words. We experiment with three different pre-processing lemmatization configurations that specifically target the stem words: (1) *Surface*, where the stem words are left as is with no further processing of the morpho-tactics that result from the segmentation of clitics; (2) *Lemma*, where the stem words are reduced to their lemma citation forms, for instance in case of verbs it is the 3rd person masculine singular perfective form; and (3) *Stem*, which is the surface form minus inflectional morphemes. It is worth noting that the Stem configuration may result in non proper Arabic words (a la IR stemming). Table illustrates examples of the three configuration schemes, with each underlined.

Features: The features we employed are of two main types: Language-independent features and Morphological features.

Language-Independent Features:

This group of features has been employed in various SSA studies for English and other European Languages.

Domain: Following Wilson (2008), we apply a feature indicating the *domain* of the document to which a sentence belongs. As mentioned earlier, each sentence has a document domain label manually associated with it. The rationale behind applying this feature is that subjective language may be distributed differently across domains. For example, based on our development data, we notice that sentences belonging to the *Sports* domain tend to be more objective than those in, for example, the *Politics* domain.

UNIQUE: Following Wiebe et al. (2004), to meaningfully incorporate a feature that accounts for the frequency of words effect, we include a *unique* feature. Namely words that occur in our corpus with an absolute frequency ≤ 5 are replaced with the token "UNIQUE".

N-GRAM: We run experiments with *N*-grams ≤ 4 and all possible combinations of them.

ADJ: For subjectivity classification, we apply a binary *has_adjective* feature indicating whether or not any of the adjectives in our manually created polarity lexicon exists in a sentence. This is motivated by Bruce and Wiebe (1999) finding that adjectives are significantly and positively correlated with subjective sentences. For sentiment classification, we employ two features, *has_POS_adjective* and *has_NEG_adjective*, each of these binary features indicate whether a POS or NEG adjective occurs in a sentence.

MSA-Morphological Features: MSA exhibits a very rich morphological system that is templatic, agglutinative, and it is based on both derivational and inflectional features. We explicitly model morphological features of *person*, *state*, *gender*, *tense*, *aspect*, and *number*. We currently do not use part of speech information explicitly in our models. We assume undiacritized text in our models.

Method: Two-stage Classification Process

In the current study, we adopt a *two-stage* classification approach. In the first stage (*Subjectivity Classification*), we build a binary classifier to sort out OBJECTIVE from SUBJECTIVE cases. For the second stage (i.e., *Sentiment Classification*) we apply binary classification that distinguishes SUBJ-POS from SUBJ-NEG cases. We disregard the neutral class of SUBJ-NEUT for our current investigation. We use an Support Vector Machine classifier SVM^{light} package (Joachims, 2008). We experiment with various kernels and parameter settings and find that linear kernels yield the best performance for our specific problem. We run experiments with *presence* vectors, i.e. for each sentence vector, the value of each dimension is binary either a 1 (regardless of how many times a feature occurs) or 0.

Experimental Conditions: We first run experiments using each of the three lemmatization settings *Surface, Lemma, Stem* using the various *N-GRAM* and *N-GRAM* combinations and then iteratively add other features exhaustively. Due to space lim-

	Surface form	**Lemma**	**Stem**
	F	F	F
	71.97	72.74	73.17
N-Gram	1g+2g+3g	1g+2g	1g+2g
Baseline	55.13	55.13	55.13

Table 4: Subjectivity Classification results on DEV data for the different lemmatization settings using N-GRAM features

itations, we present here the most interesting feature combinations namely: N-GRAM for all lemmatization settings, Stem+Morph, Lemma+ {DOMAIN, ADJ, UNIQUE}, Stem+Morph+{DOMAIN, ADJ, UNIQUE}.

With all the three settings, clitics that are split off words are kept as separate features in the sentence vectors.

Results and Evaluation

We divide our data into 80% for 5-fold cross-validation and 20% for test. For experiments on the test data, the 80% are used as training data. We have two settings, a development setting (DEV) and a test setting (TEST). In the development setting, we run the typical 5 fold cross validation where we train on 4 folds and test on the 5th and then average the results. In the test setting, we only ran with the best configurations yielded from the DEV conditions. In TEST mode, we still train with 4 folds but we test on the test data exclusively, averaging across the different training folds but crucially maintaining the same proportion of 4 folds for training. It is worth noting that the test data is larger than any given dev data (20% of the overall data set for test, vs. 16% for any DEV fold). We report results using *F*-measure (*F*).

Subjectivity

Table shows the overall results obtained in the DEV settings using N-GRAM features only. We experimented exhaustively with the different n-gram sizes and their combinations. As indicated in the table, different N-GRAM feature settings resulted in the best results for the different lemmatization configuration. In the Surface configuration, 1g+2g+3g N-GRAM feature combination yields the best results for this setting, while 1g+2g yields the best results for Lemma and Stem respectively. Stem yields the best classification results overall. All three results for Surface, Lemma, and Stem respectively outperform the Baseline which is a majority class baseline.

Furthermore, Table shows the impact of adding morphological features to the best yielding basic lemmatization setting Stem, as illustrated in the column Stem+Morph. Stem+Morph with N-GRAM feature 1g+2g+3g yields 73.48 *F* which is an increase over Stem without explicit morphological features that yields 73.17% *F*.

Adding language-independent features, illustrated in Table , the *ADJ* feature does not help in either the *Lemma* nor in the *Stem+Morph* settings, however, it helps

	Stem	Stem+Morph
	73.17	73.48
N-Gram	1g+2g	1g+2g+3

Table 5: Subjectivity Classification results on DEV data for Stem and Stem+Morph

	BASE	+ADJ	+DOMAIN	+UNIQUE
Lemma	72.74	72.12	72.64	73.05
Stem	73.17	73.22	72.78	**72.85**
Stem+Morph	**73.48**	**73.42**	**73.55**	72.53

Table 6: Subjectivity Classification results on DEV data for Lemma, Stem, and Stem+Morph +language-independent features

with the *Stem*. The *DOMAIN* feature improves the results only with the *Stem+Morph*. In addition, the *UNIQUE* feature modestly helps classification in the *Lemma* setting, but has a negative impact on both the *Stem* and the *Stem+Morph* settings. Table shows that although performance on TEST set drops with some settings on *Stem+Morph*, 6.25% improvement of *F* is acquired by applying the *ADJ* feature.

Sentiment

As Table shows, similar to the subjectivity results, the *Stem* setting performs better than the other two lemmatization settings, with 56.87% *F*, compared to 52.53% *F* for the *Surface* and 55.01% *F* for the *Lemma*, although this is still outperformed by the majority class baseline. Again, adding the morhology-based features helps improve the classification: The *Stem+Morph* is better than the *Stem* by about 1.00% *F*, as shown in Table . Table shows that whereas adding the *DOMAIN* feature helps in the *Lemma*, emphStem, and *Stem+Morph* settings, the *UNIQUE* feature only improves classification with the *Stem+Morph*. Adding the *ADJECTIVE* feature improves performance significantly: Improvements of 33.71% *F* for the *Lemma* setting, 34.06% *F* for the *Stem*, and 33.09% *F* for the *Stem+Morph* are possible. As Table shows, while performance on TEST data drops with application of the the *UNIQUE* feature, it slightly improves when the *DOMAIN* feature is added and significantly when the *ADJ* feature is used (the latter reaching 95.52% *F*).

Error Analysis

We performed an analysis of the errors made by the system in both the subjectivity and sentimment cases.

Stem+Morph	+ADJ	+DOMAIN	+UNIQUE
65.29	**71.54**	**63.54**	**65.17**

Table 7: Subjectivity Classification results on TEST data for Stem+Morph+language-independent features

	Surface form	Lemma	Stem
	F	F	F
	52.53	55.01	56.87
N-Gram	1g	1g	1g
Baseline	**58.65**	**58.65**	**58.65**

Table 8: Sentiment Classification results on DEV data for the different lemmatization settings using N-GRAM features

	Stem	Stem+Morph
	56.87	57.84
N-Gram	1g	1g

Table 9: Sentiment Classification results on DEV data for Stem and Stem+Morph

	BASE	+ADJ	+DOMAIN	+UNIQUE
Lemma	55.01	88.72	57.21	54.22
Stem	56.87	90.93	57.66	55.55
Stem+Morph	57.84	**90.93**	58.03	58.22

Table 10: Sentiment Classification results on DEV data for Lemma, Stem, and Stem+Morph +language-independent features

Stem+Morph	+ADJ	+DOMAIN	+UNIQUE
52.12	**95.52**	53.21	51.92

Table 11: Sentiment Classification results on TEST data for Stem+Morph+language-independent features

Error Analysis of Subjectivity Classification

The errors made by the system in the case of subjectivity analysis show that subjectivity is very context sensitive. Example 15 below, which was falsely classified by the system as subjective, illustrates that subjective words (e.g., “Alm‘tqlyn” “the detainees”, “AstslmwA” “they surrendered”, “’frĝt” “it released”, and “Astd‘thm” “it issued summons to them”) frequently occur in objective sentences and hence potentially confuse the classifier.

(١٥) وذكر مصدر في الشرطة أن أربعة عشر من المعتقلين السابقين ال ٣٥ في
حركة حماس الذين أفرجت عنهم السلطة الفلسطينية استسلموا للشرطة الفلسطينية
التي كانت استدعتهم.

English: A police source said that fourteen of the 35 former detainees in the Hamas movement, released by the Palestinian Authority, surrendered to Palestinian police, which had issued summons.
Buckwalter: w*kr mSdr fy Al$rTp In OrbEp E$r mn AlmEtqlyn AlsAbqyn Al 35 fy Hrkp HmAs Al*yn Ofrjt Enhm AlslTp AlflsTynyp AstslmwA ll$rTp AlflsTynyp Alty kAnt AstdEthm.

The system also failed to classify some of the sentences with time-constrained propositions (e.g., those where items like “ytwq‘” “it is expected”) occur and hence shift the class from factual to hypothetical/expected. Example 16 illustrates this specific case along with the fact that not only adjectives but polarized nouns (e.g., “mwAĝhAt” “confrontations” “qtlY” “killed people”) as well can contribute to shifting a sentence class. Based on these specific observations, it might help to apply a feature indicating whether items capable of conditioning the time scope of propositions (e.g., “expected”, “supposed”) exist or not. Expanding the polarity lexicon beyond adjectives is also desirable.

(١٦) ويتوقع أن يلتقي الزعيمان الاسرائيلي إيهود باراك والفلسطيني ياسر عرفات
غدا الاثنين في شرم الشيخ للتوصل إلى اتفاق حول وقف المواجهات التي أدت
إلى سقوط ١٠٦ قتلى معظمهم من الفلسطينيين وأكثر من ثلاثة آلاف جريح.

English: The Israeli leader Ehud Barak and the Palestinian leader Yasser Arafat are expected to meet on Monday in Sharm el-Sheikh to reach agreement on ending the confrontations that led to the fall of 106 deaths, most of them Palestinians, and more than three thousand injured.
Buckwalter: wytwqE On yltqy AlzEymAn AlIsrA}yly Iyhwd bArAk wAlflsTyny yAsr ErfAt gdA AlAvnyn fy $rm Al$yx lltwSl IlY AtfAq Hwl wqf AlmwAjhAt Alty Odt IlY sqwT 106 qtlY mEZmhm mn AlflsTynyyn wOkvr mn vlAvp *vert*lAf jryH.

We also observed that most of the sentences belonging to the SPORT domain in our development set are objective. The system seemed to associate words from the sports domain more with the objective class and hence incidentally misclassifies some sentences, as in example 17 below. Example 17 was misclassified as objective, even though it has polarized words. Observably, the example has more than one polarized items (e.g., “sĝlh hAfl” “track record”, “’ahdr” “wasted”, “_tmynp” “precious”, “mrtfʿ” “high”). It may thus prove useful to add a feature related to the number of polarized words used in a sentence.

(١٧) كثرت الترشيحات لمنتخبي الكويت وكوريا الجنوبية للمنافسة على لقب
البطولة نظرا لسجلهما الحافل في البطولات السابقة، الأول لم يظهر بمستواه المعهود
وأهدر نقطتين ثمينتين أمام إندونيسيا، والثاني كان مستواه مرتفعا جدا.

English: There have been many nominations for the national teams of Kuwait and South Korea to compete for the championship title because of of their track records in previous championships; the first did not keep up to its usual high level and wasted two valuable points against Indonesia, and the second showed a very high level.
Buckwalter: kvrt Altr$yHAt lmntxby Alkwyt wkwryA Aljnwbyp llmnAfsp ElY lqb AlbTwlp nZrA lsjlhmA AlHAfl fy AlbTwlAt AlsAbqp , AlOwl lm yZhr bmstwAh AlmEhwd wOhdr nqTtyn vmyntyn OmAm IndwnysyA, wAlvAny kAn mstwAh mrtfEA jdA.

Even though objective examples were observed to include numbers and reporting verbs more than subjective sentences, polarized content still co-existed with this objectivity cues (as is the case of example 18 below, which was misclassified as objective). A careful consideration of example 4 suggests that strongly polarized words (e.g., “mĝAzr” “massacres”) can act as strong subjectivity cues. Our system does not currently have access to the strength of polarized words. Adding stength values to the polarity lexicon may improve the system’s performance.

(١٨) وقد قتل أكثر من ٩٠ شخصا منذ بداية تشرين الأول أكتوبر في مجازر
واعتداءات نسبت إلى الجماعات الاسلامية المسلحة المعادية لسياسة الوفاق الوطني
التي يتهجها الرئّيس عبد العزيز بوتفليقة وفق حصيلة استنادا إلى الصحف.

English: More than 90 people has been killed since the beginning of October in the massacres and attacks attributed to armed Islamist groups hostile to the national reconciliation policy pursued by President Abdelaziz Bouteflika, according to an outcome based on the newspapers.
Buckwalter: wqd qtl Okvr mn 90 $xSA mn* bdAyp t$ryn AlOwl Oktwbr fy mjAzr wAEtdA’At nsbt IlY AljmAEAt AlIslAmyp AlmslHp AlmEAdyp lsyAsp AlwfAq AlwTny Alty ynthjhA Alr}ys Ebd AlEzyz bwtflyqp wfq HSylp AstnAdA IlY Al-SHf.

Error Analysis of Sentiment Classification

The system misclassified some sentences whose adjective polarities are shift ed as a result of surrounding negation or polarity shifters (Polanyi and Zaenen, 2006).

Example 19 below has the adjective “dblwmAsyp” “diplomatic” preceded by the polarity shifter “qTE” “cutting”, and was misclassified as positive. Applying a negation feature may thus improve the system’s performance.

(١٩) وأضاف "\أول ترجمة لمثل هذا الدعم لا تكون إلا بقطع العلاقات
الدبلوماسية وكل أشكال التبادل مع إسرائيل ومقاطعة إسرائيل على أوسع نطاق
اقتصاديا "\.

English: He added, “[T]he first translation of such support can only happen by cutting off diplomatic relations and all forms of exchange with Israel ,and economically boycotting at on the scale”.
Buckwalter: wODAf “Owl trjmp lmvl h‘*A AldEm lA tkwn IlA bqTE AlElAqAt AldblwmAsyp wkl O$kAl AltbAdl mE IsrA}yl wmqATEp IsrA}yl ElY OwsE nTAq {qtSAdyA”.

The system also did not cater for another category of polarity shifters (Polanyi and Zaenen, 2006), i.e., *epistemic modality* (with *hedges* like “rbmA” “perhaps” and *boosters* like “bAltOkyd” “certainly”). This resulted in errors like example 20 below where the hedging phrase “Al.hd mn” “reducing” softens the claim and hence alters the polarity.

(٢٠) وجاء في بيان نشر في ختام إجتماع لحّكومة أن "\رئيس الوزراء أكد أمام
الحكومة إن قمة شرم الشيخ تهدف إلى وقف العنف واقامة هيئة تعنى بالحد من
مخاطر تجدد العنف ودراسة الأحداث التي وقعت منذ أسبوعين "\.

English: According to a statement released at the conclusion of a meeting of the government, “[T]he prime minister stressed to the government that the Sharm El-Sheikh summit aimed at halting the violence and establishing a body to reduce the risk of renewed violence and studying the events that took place two weeks ago.”
Buckwalter: wjA’ fy byAn n$r fy xtAm IjtmAE llHkwmp On {r}ys AlwzrA’ Okd OmAm AlHkwmp In qmp $rm Al$yx thdf IlY wqf AlEnf wIqAmp hy}p tEnY bAlHd mn mxATr tjdd AlEnf wdrAsp AlOHdAv Alty wqEt mn* OsbwEyn”.

The system also misclassified sentences with adjectives that were labeled neutral in the polarity lexicon, but which are polarized from the perspective (Lin et al., 2006) of the witer or person quoted in a news story. Example 21 below was missclassified as negative, although the adjective “Muslim” is positive from the perspective of the qouted person. Perspective identification may help assign polarities to certain adjectives.

(٢١) وأكد إن إيران "\ستساعد إقتصاديا شعب العراق المسلم "\.

English: He stressed that Iran would “help the Muslim people of Iraq economically.”
Buckwalter: wOkd In IyrAn “stsAEd IqtSAdyA $Eb AlErAq Almslm”.

The system also misclassified some sentences where multi-word polarized expressions (MWE) were used. Example 22 has an occurrence of a negative MWE (i.e., “Alal‘b bAlnAr” “playing with fire”), but was misclassified as positive. This problem may be solvable by expanding the polarity lexicon with MWE.

(٢٢) وأخذ بيريز على الفلسطينّين "\اللعب بالنار "\، وقال "\يجب أن يفهم
الفلسطينيون أنهم يلعبون بالنار، ليس فقط مع إسرائيل بل مع العالم كله "\.

English: Peres balmed the Palestinians to for “playing with fire” and said, “Palestinians must understand that they are playing with fire, not only with Israel but with the whole world.”
Buckwalter: wOx* byryz ElY AlflsTynyyn {AllEb bAlnAr}, wqAl {yjb On yfhm AlflsTynywn Onhm ylEbwn bAlnAr, lys fqT mE IsrA}yl bl mE AlEAlm klh”n.

Related Work

Regarding annotation of data for SSA, work on the news genre is most relevant to us. Wiebe, Wilson and Cardie (2005) describe a fine-grained news corpus manually labeled for SSA at the word and phrase levels. Their annotation scheme involves identifying the *source* and *target* of sentiment as well as other related properties (e.g., the *intensity* of expressed sentiment). Our work is less fine grained on the one hand, but we label our data for domain as well as subjectivity.

Balahur et al. (2009) report work on labeling quotations from the news involving one person mentioning another entity and maintain that quotations typically contain more sentiment expressions than other parts of news articles. Our work is different from that of Balahur et al. (2009) in that we label all sentences regardless whether they include quotations or not. Balahur et al. (2009) found that entities mentioned in quotations are not necessarily the target of the sentiment, and hence we believe that SSA systems built for news are better if they focus on all the sentences of articles rather than quotations alone (since the target of sentiment may be outside the scope of a quotation, but within that of the sentence to which a quotation belongs).

As for SSA systems, several sentence- and phrase-level classifiers have been built, (e.g., Wiebe, Bruce and O’Hara, 1999; Yi et al., 2003; Yu and Hatzivassiloglou, 2003; Kim and Hovy, 2004). Yi et al. (2003) present an NLP-based system that detects all references to a given subject, and determines sentiment in each of the references. Similar to Yi et al. (2003), Kim and Hovy (2004) present a sentence-level system that, given a topic detects sentiment towards it. Our approach differs from both Yi et al. (2003) and Kim and Hovy (2004) in that we do not detect sentiment toward specific topics. Also, we make use of *N*-gram features beyond unigrams and employ elaborate *N*-gram combinations.

Yu and Hatzivassiloglou (2003) build a document- and sentence-level subjectivity classification system using various *N*-gram-based features and a polarity lexicon. They report about 97% F-measure on documents and about 91% F-measure on sentences from the *Wall Street Journal* (WSJ) corpus. Some of our features are similar to those used by Yu and Hatzivassiloglou (2003), but we exploit additional features. Wiebe, Bruce and O'Hara (1999) train a sentence-level probabilistic classifier on data from the WSJ to identify subjectivity in these sentences. They use POS features, lexical features, and a paragraph feature and obtain an average accuracy on subjectivity tagging of 72.17%. Again, our feature set is richer than Wiebe, Bruce and O'Hara (1999).

The only work on Arabic SSA we are aware of is that of Abbasi, Chen and Salem (2008). They use an entropy weighted genetic algorithm (EWGA) for both English and Arabic Web forums at the document level. They exploit both syntactic and stylistic features. Abbasi et al. use a root extraction algorithm and do not use morphological features. They report 93.6% accuracy. Their system is not directly comparable to ours due to the difference in data sets and tagging granularity.

Conclusion

In this paper, we present a novel annotation layer of SSA to an already labeled MSA data set, the PATB Part 1 ver. 3.0. To the best of our knowledge, this layer of annotation is the first of its kind on MSA data of the newswire genre. We will make that collection available to the community at large. We motivate SSA for news and summarize our linguistics-motivated guidelines for data annotation and provide examples from our data set. We also build a sentence-level SSA system for MSA contrasting language independent only features vs. combining language independent and language-specific feature sets, namely morphological features specific to Arabic. We also investigate the level of stemming required for the task. We show that the *Stem* lemmatization setting outperforms both *Surface* and *Lemma* settings for the SSA task. We illustrate empirically that adding language specific features for MRL yields improved performance. Similar to previous studies of SSA for other languages, we show that exploiting a polarity lexicon has the largest impact on performance. We also identify several areas where the system makes errors, with a view to future improvement.

References

Abbasi, A., Chen, H., Salem, A. (2008): Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums. *ACM Trans. Inf. Syst.*, Vol. 26, pp. 1–34.

Abdul-Mageed, M. (2008): Online news sites and Journalism 2.0: Reader comments on Al Jazeera Arabic. *Triple C: Cognition, Communication, Cooperation*, Vol. 6, No. 2, pp. 59–76.

Abdul-Mageed, M., Diab, M. (2011): Subjectivity and Sentiment Annotation of Modern Standard Arabic Newswire. In: *LAW V: Fifth Linguitic Annotation Workshop : Proceedings of the Workshop.* Stroudsburg (Philadelphia, USA) : Association for Computational Linguistics, pp. 110–118. Available at: http://www.aclweb.org/anthology/W11-0413.

Bach, K., Harnish, R. M. (1979): *Linguistic communication and speech acts.* Cambridge (MA) : MIT Press, 327 pp.

Balahur, A., Steinberger, R. (2009): Rethinking Sentiment Analysis in the News: from Theory to Practice and Back. In: *WOMSA '09 : Proceedings of the 1st Workshop on Opinion Mining and Sentiment Analysis.* Sevilla : University of Sevilla, pp. n/a. Available at: http://130.203.133.150/viewdoc/versions?doi=10.1.1.157.3642.

Balahur, A., Steinberger, R., van der Goot, E., Pouliquen, B., Kabadjov, M. (2009): Opinion Mining on Newspaper Quotations. In: *2009 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology.* IEEE Computer Society, pp. 523–526. (ISBN 978-0-7695-3801-3.)

Banfield, A. (1982): *Unspeakable Sentences: Narration and Representation in the Language of Fiction.* Boston : Routledge & Kegan Paul, 340 pp.

Bruce, R., Wiebe, J. (1999): Recognizing subjectivity. A case study of manual tagging. *Natural Language Engineering*, Vol. 5, No. 2, pp. 187–205.

Diab, M., Hacioglu, K., Jurafsky, D. (2007): Automatic processing of Modern Standard Arabic text. In: *Arabic Computational Morphology.* Heidelberg : Springer, pp. 159–179. (ISBN 978-1-4020-6046-5.)

Fowler, R. (1991): *Language in the News: Discourse and Ideology in the Press.* London : Routledge, 254 pp. ISBN 0-415-01418-2.

Habash, N., Rambow, O., Roth, R. (2009): MADA+TOKAN: A Toolkit for Arabic Tokenization, Diacritization, Morphological Disambiguation, POS Tagging, Stemming and Lemmatization. In: Choukri, Kh., Maegaard, B. *Proceedings of the Second International Conference on Arabic Language Resources and Tools* [MEDAR '09]. [s. l.] : MEDAR Consortium, pp. 102–109. (ISBN 2-9517408-5-9.)

Chouliaraki, L., Fairclough, N. (1999): *Discourse in late modernity : Rethinking critical discourse analysis.* Edinburgh : Edinburgh University Press, 168 pp. ISBN 0-7486-1082-0.

Joachims, T. (2008): SVM^{light} : Support Vector Machine v. 6.02 [software].

Bach, K., Harnish, R. M. (1979): *Linguistic communication and speech acts.* Cambridge (MA) : MIT Press, 327 pp.

Kim, S., Hovy, E. (2004): Determining the sentiment of opinions. In: *Coling 2004 : 20th International Conference on Computational Linguistics : Proceedings of the Conference.* Stroudsburg (Philadelphia, USA) : Association for Computational Linguistics, pp. 1367–1373.

Lin, W. H., Wilson, T., Wiebe, J., Hauptmann, A. (2006): Which side are you on?: identifying perspectives at the document and sentence levels. In: *Proceedings of the Tenth Conference on Computational Natural Language Learning.* Stroudsburg (Philadelphia, USA) : Association for Computational Linguistics, pp. 109–116.

Maamouri, M., Bies, A., Buckwalter, T., Mekki, W. (2004): The penn arabic treebank: Building a large-scale annotated arabic corpus. In: *NEMLAR Conference on Arabic Language Resources and Tools.* , pp. 102–109.

Palmer, F. (1986): *Mood and Modality. 1986.* Cambridge : Cambridge University Press, x+243 pp.

Polanyi, L., Zaenen, A. (2006): Contextual valence shifters. In: *Computing attitude and affect in text: Theory and applications.* Heidelberg : Springer, pp. 1–10.

Quirk, R., Greenbaum, S., Close, R. A., Quirk, R. (1974): *A university grammar of English.* London : Longman, 484 pp.

Searle, J. R. (1975): A taxonomy of speech acts. In: Gunderson, K. *Language, mind, and knowledge.* Minneapolis : University of Minnesota Press, pp. 344–369.

Tsarfaty, R., Seddah, D., Goldberg, Y., Kuebler, S., Versley, Y., Candito, M., Foster, J., Rehbein, I., Tounsi, L. (2010): Statistical Parsing of Morphologically Rich Languages (SPMRL) What, How and Whither. In: *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages.* Stroudsburg (Philadelphia, USA) : Association for Computational Linguistics, pp. 1–12.

Van Dijk, T. A. (1988): *News as discourse.* Mahwah (NJ) : Lawrence Erlbaum Associates, 200 pp.

Wiebe, J. (1994): Tracking point of view in narrative. *Computional Linguistics*, Vol. 20 No. 2, pp. 233–287.

Wiebe, J., Bruce, R., O'Hara, T. (1999): Development and use of a gold standard data set for subjectivity classifications. In: *37th Annual Meeting of the Association for Computational Linguistics : Proceedings of the Conference.* Stroudsburg (Philadelphia, USA) : Association for Computational Linguistics, pp. 246–253. (ISBN 1-55860-609-2.)

Wiebe, J., Wilson, T., Bruce, R., Bell, M., Martin, M. (2004): Learning subjective language. *Computational Linguistics*, Vol. 30, No. 3, pp. 277–308. (ISSN 0891-2017)

Wiebe, J., Wilson, T., Cardie, C. (2005): Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, Vol. 39, No. 2, pp. 165–210. (ISSN 1574-020X)

Wilson, T., Wiebe, J., Hoffmann, P. (2009): Recognizing Contextual Polarity: an exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, Vol. 35, No. 3, pp. 399–433. (ISSN 0891-2017)

Wodak, R., Meyer, M. (2009): *Methods of critical discourse analysis.* London : Sage, pp. 1–33. Chapter 1: Critical discourse analysis: History, agenda, theory and methodology.

Yi, J., Nasukawa, T., Bunescu, R., Niblack, W. (2003): Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In: *Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM 2003).* IEEE Computer Society, pp. 427–434. (ISBN 0-7695-1978-4.)

Yu, H., Hatzivassiloglou, V. (2003): Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In: *Proceedings of the 2003 conference on Empirical methods in natural language processing. Vol. 10 (EMNLP '03).* Stroudsburg (Philadelphia, USA) : Association for Computational Linguistics, pp. 129–136.

The smallest, cheapest, and best: Superlatives in Opinion Mining

Silke Scheible

*School of Languages, Linguistics and Cultures, University of Manchester, Oxford Road, Manchester United Kingdom, **

e-mail: `Silke.Scheible@manchester.ac.uk`

Abstract

This article introduces superlatives as special indicators for product features in customer reviews. The investigation shows that one type of superlative (called 'ISA') is of particular relevance, as instances in this class tend to contain both a feature string and its associated opinion word. An identification of the components of such superlative comparisons can therefore help to solve two Opinion Mining tasks at once: Feature and Opinion Word Identification. The study further introduces and evaluates a novel tool that can reliably identify such superlatives, and extract from them potential product feature strings and opinion words.

Key words

superlatives; opinion mining; product features; opinion words; adjectives

Introduction

In recent years, the domain of product reviews has attracted much attention in the area of Sentiment Analysis and Opinion Mining. While the main goal of the former is classification of documents, sentences, phrases or words as positive or negative, the interest in Opinion Mining lies in extracting information about which entities or features of entities are considered as positive or negative, and to summarise this information (Hu and Liu, 2004; Popescu and Etzioni, 2005; Carenini et al., 2005). This is of great benefit not only for companies who want information about customer's opinions on their products, but also for recommendation systems whose purpose is to assist customers in deciding which product to buy. In general, Opinion Mining systems are required to solve the following main tasks (e.g. Hu and Liu, 2004):

1. Feature Identification
2. Opinion Word Identification
3. Sentiment Classification
4. Opinion Summarisation

[1] This paper is a printed version of the paper published in the Proceedings of WASSA-2010, see `http://gplsi.dlsi.ua.es/congresos/wassa2010/fitxers/WASSA2010_Proceedings_.pdf`.

The first step is to identify features of the products that customers are interested in, usually by using data mining and natural language processing techniques. Hu and Liu (2004) define the term "product feature" as representing both components of an object (e.g. zoom) and their respective attributes (e.g. size).[2] The next step is to identify sentences in the reviews that express opinion s about these features. This involves distinguishing opinion words from factual words (subjectivity recognition). To address (3), the system has to determine whether a statement of opinion is positive or negative. Finally, the system also requires techniques for summarising this information (Carenini et al., 2005; Carenini and Cheung, 2008).

So far, none of the studies in Sentiment Analysis or Opinion Mining have specifically looked at the role of superlatives in these areas. While it has been generally acknowledged that there is a positive correlation between subjectivity and the use of *adjectives* (e.g. Hatzivassiloglou and Wiebe, 2000), there has not yet been a thorough investigation of *superlative* adjectives and adverbs in this context. This article aims to show that some types of superlative represent a special linguistic means of expressing opinions about products. Consider for example:

(1) The Panasonic TC-P54G10 is the *best* plasma TV on the market.
(2) It has the *clearest* picture I have ever seen.

I claim that superlative constructions like (1) and (2) act as special indicators of product features, which contain both the opinion word (the superlative, italicised) and the feature string (underlined). This means that the identification of the components of such superlative comparisons addresses two Opinion Mining tasks at once: Feature and Opinion Word Identification. This article provides evidence for this claim, and introduces a tool which can be used to reliably identify superlatives of interest and extract potential product feature strings from them.

Previous approaches

Existing work on identifying product features (Task 1) often relies on the simple heuristic that explicit features are expressed as noun phrases. While this narrows down the set of product feature candidates, it is clear that not all noun phrases represent product features. Various approaches to further limiting this set have been proposed. The two most notable ones are Hu and Liu (2004) and Popescu and Etzioni (2005).

Hu and Liu (2004) suggest that nouns or noun phrases that occur frequently in reviews for a particular product are likely to be features. To identify frequent features they use association mining, and then apply heuristic-guided pruning to further refine their results. They further assume that adjectives appearing in the same sentence as frequent features are opinion words, thereby solving Task 2 (however, at the cost of precision). In addition, retrieving nouns and noun phrases that co-occur with these opinion words in other sentences helps their system to identify so-called *infrequent* features, which are also of great interest (Pang and Lee, 2008).

Popescu and Etzioni (2005), on the other hand, consider product features to be concepts that stand in particular semantic relationships with the product (for

[2] In this definition, the object itself is also a feature.

example, a camera may have "properties" size, weight, etc., while the lens, flash, etc. stand in a "part" relationship with the camera). Their strategy for identifying such features is to search for corresponding meronymy discriminators. This approach achieves better performance than the one employed by Hu and Liu (2004), but no sentiment analysis is carried out, and opinion words have to be identified in a second step.

Although a previous study by Jindal and Liu (2006) investigated graded adjectives in the context of customer reviews, their study is not suitable for identifying product features. They investigate the topic of comparative sentence mining, whose goal is to identify sentences in evaluative texts on the web that express "an ordering relation between two sets of entities with respect to some common features", and to extract comparative relations from the identified sentences. A follow up study by Ganapathibhotla and Liu (2008) builds on these findings and aims to determine which of the extracted entities in a comparison are preferred by its author. However, as Jindal and Liu (2006) apply their vector approach to *every* graded adjective in the corpus, this involves a large amount of cases which do not modify "product features" (as identified and annotated by Hu and Liu (2004) in the same corpus). As a consequence, their system is not suitable for the task of identifying product features. Furthermore, even though Jindal and Liu's system aims to identify the components of superlative comparisons, a closer study showed that their approach does not distinguish between different types of superlatives, leading to incorrect analyses of superlative constructions (Scheible, 2007). The current study takes different superlative surface constructions into account, and suggests that a particular subclass of superlatives (namely, 'ISA superlatives') is especially useful in identifying product features.

Superlatives in Opinion Mining

Superlatives describe a well-defined class of word forms which (in English) are derived from adjectives or adverbs in two different ways: Inflectionally, where the suffix *-est* is appended to the base form of the adjective or adverb (e.g. *lowest*, *nicest*, *smartest*), or analytically, where the base adjective/adverb is preceded by the markers *most/least* (e.g. *most interesting*, *least beautiful*). Certain adjectives and adverbs have irregular superlative forms: *good (best)*, *bad (worst)*, *far (furthest/farthest)*, *well (best)*, *badly (worst)*, *much (most)*, and *little (least)*. In linguistics, superlatives are usually introduced alongside comparatives as special forms of adjectives or adverbs which are used to compare two or more things, as for example in:

(3) Bill is *taller* than Sue. [comparative]

(4) {Joe} is the *tallest* [boy at school]. [superlative]

Superlative constructions like (4) express a comparison between a target entity T (Joe; curly brackets) and its comparison set CS (the other boys at school; square brackets). An investigation of superlative forms showed that two types of relation hold between a superlative target and its comparison set (Scheible, 2007):

Relation 1: Superlative relation

Relation 2: IS-A relation (hypernymy)

The superlative relation specifies a property which all members of the set share, but which the target has the highest (or lowest) degree or value of. The IS-A relation expresses the membership of the target in the comparison class (e.g. its parent class in a generalisation hierarchy). For example, in (4), the superlative relation implicitly specifies the property *height*, which applies to all members of the comparison set *boys at school*. Of this set, the target *Joe* has the greatest *height* value. The IS-A relation states that *Joe* is a member of the set *boys at school*. Both relations are of great interest for relation extraction, and Scheible (2009) discusses their use in applications such as Question Answering (QA) and Ontology Learning. Superlatives occur in a variety of syntactic structures which usually represent different types of comparisons. Scheible (2009) developed a classification of superlatives based on surface forms (illustrated in Table 1).

Table1: Superlative classes

	Class	Example
(a)	ISA	ISA-1: The Panasonic TC-P54G10 is the best plasma TV on the market.
(b)		ISA-2: The Samsung is considered the most stylish plasma TV.
(c)	DEF	I bought the cheapest plasma TV.
(d)	INDEF	Plasma TVs represent a most compelling option for home entertainment.
(e)	FREE	The 37” size is best when you are 8–10 feet away from the screen.
(f)	ADV	HD TVs most commonly use progressive scan for 1280×720.
(g)	IDIOM	The 42PC1RR won the Best Plasma TV Award this year.
(h)	PP	The TV weighs about 57 pounds at most.
(i)	PROP	Most cheap TVs have poor quality scalers.

Superlatives belonging to the ISA class are incorporated in a definite NP and contain a clear-cut comparison between a target item and its comparison set. In example (a) in Table 1, the Panasonic TC-P54G10 is compared to other plasma TVs on the market with respect to its overall quality. The difference between the ISA-1 and ISA-2 subclasses lies in the way in which the relation between target and comparison set is expressed. In the case of ISA-1 superlatives, the verb “to be” or appositive form is used, while ISA-2 superlatives involve other forms (e.g. other copula verbs). While superlatives classified as DEF are also incorporated in a definite NP, they differ from members of the ISA class in that the target of comparison is not independently specified in the context. In example (c) the comparison remains ‘implicit’ as the target is not specified in the sentence, except as that which satisfies the superlative NP. When superlative forms are incorporated in an indefinite NP they are classified as INDEF (d). Members of this class are often used as intensifiers. In the FREE class, on the other hand, superlative forms are not incorporated in a noun phrase but occur freely in the sentence. This often makes the comparison less easy to pinpoint: (e) does not compare the 37” size with other screen sizes, but rather the quality of the 37” size viewed from different locations in the room. Superlatives that are derived from adverbs form their own class, ADV (f). Finally, the IDIOM, PP, and PROP classes contain superlatives which do not express proper comparisons: IDIOM contains superlatives that occur as part of an idiom (g), PP contains so-called PP superlative constructions (h), and PROP includes uses of *most* as a proportional quantifier (i).

This study argues that superlatives of the type ISA are of particular importance in Opinion Mining as they make explicit the IS-A relation that holds between target and comparison set (cf. Relation 2 above). This means that both their target and comparison set are explicitly realised in the text, where the target string often expresses the product, the CS string expresses a feature while the superlative itself expresses the opinion word (as in (a) and (b)). The present study rests on the following claims:

1. ISA superlatives are special indicators for sentences containing product features.
2. The product feature usually appears within their T or CS string, while the superlative expresses its respective opinion word.

The next section briefly describes the data used to support these claims.

Data

The investigation described in this article uses Hu and Liu's corpus of customer reviews, which was not only the basis of their own study of opinion feature mining Hu and Liu (2004), but has been used as test set by other studies as well, e.g. Popescu and Etzioni (2005). The corpus contains reviews of five products: two digital cameras (Canon G3 and Nikon Coolpix 4300), one mobile phone (Nokia 6610), an mp3 player (Creative Labs Nomad Jukebox Zen Xtra 40GB), and a dvd player (Apex AD2600 Progressive-scan)[3] Sentences in this corpus have been manually annotated with information about product features. Each feature is taken to express an opinion, and labelled as *positive* or *negative* in terms of values on a six-point scale, where [+3] and [+1] stand for the strongest positive and weakest positive opinions, respectively, and [−3] and [−1] stand for the strongest and weakest negative opinions. Hu and Liu's corpus contains 4,259 sentences altogether, of which 1,728 include at least one product feature (40.6%). The remaining sentences in the corpus either contain no product feature (2,217 altogether, 52.1%), or describe a review title, in which case they have been excluded from consideration (314 instances, 7.4%). The corpus contains a total of 230 superlatives in 4,259 sentences, which means that there is around one superlative in every 18 sentences. All 230 superlatives found in the corpus were annotated with class labels as shown in Table 1.

ISA-Superlatives as product feature indicators

This section aims to provide support for the claim that superlatives are special indicators of product features in customer reviews. In particular, I will show that this especially applies to a subgroup of superlatives (ISA) by analysing the distribution of feature labels across the eight superlative classes in Hu and Liu's corpus of customer reviews. Table 2 shows the overall distribution of superlative classes in

[3] `http://www.cs.uic.edu/~liub/FBS/CustomerReviewData.zip`

Table2: Distribution of features

Class	#S	#T	#F	#N
ISA	71	2	53	16
DEF	45	9	16	20
INDEF	15	10	3	2
FREE	3	1	2	0
ADV	10	0	4	6
IDIOM	12	0	5	7
PP	27	1	13	13
PROP	47	0	23	24
TOTAL	230	23	119	88
	100%	10%	51.7%	38.3%

the corpus (columns 1 and 2). The ISA class is the most frequent with 71 instances (30.9%) (of which ISA-1 accounts for 63 instances, and ISA-2 for 8). The table further shows the proportion of title sentences (T), feature-containing sentences (F), and non-feature containing sentences (N) among the 230 superlative-containing sentences (S) in the corpus. The last row (TOTAL) indicates that the proportion of feature-containing sentences among them is higher (at 51.7%) than the average for all sentences (which is 40.6%, cf. Section 4). What is especially striking is that features are particularly highly represented among sentences containing ISA superlatives: Of 71 ISA superlatives in the data set, 53 occur in a sentence involving a feature (74.6 %). This suggests that membership in the ISA class is a good indicator of the sentence containing a product feature.

A closer investigation of the data reveals further interesting results. Among the 119 superlative sentences that contain a feature (column "F"), not all superlatives directly contribute to the evaluation of the feature. For example, the superlative "most" in (5), which belongs to the PROP class, is not directly involved in the evaluation of the feature "firewire" as [−1]. In contrast, the ISA superlative "best" in (6) is directly responsible for the positive [+3] rating of the feature "dvd player".

(5) it does n't have firewire, not a real complaint since most windows users do n't generally have firewire cards themselves . [Creative]

(6) i think , apex is the *best* [dvd player you can get for the price] . [Apex]

An assessment of all feature-containing sentences with respect to the involvement of the superlative in the feature-rating shows that the IDIOM, PP, and PROP classes are of little relevance, while ISA-1 and ISA-2 clearly are, with the superlative form acting as opinion word evaluating the feature, or acting as intensifier of an opinion word, as for example "complaint" in (7).

(7) [my] biggest [complaint] is the battery life or lack there of . [Creative]

Furthermore, in 34 out of the 46 feature-containing ISA-1 instances (73.9%) and in 6 out of 7 ISA-2 instances (85.7%), the feature is a substring of either the target (as shown in (7)) or the comparison set spans (6). The importance of the

ISA class is further supported by an investigation which showed that Hu and Liu's annotation is not always consistent. Several of the 16 ISA-1 instances that did not receive a feature label in Hu and Lu's annotation (column "N" in Table 2) do in fact modify a feature. For example, (8) and (9) make a similar positive statement about a camera, however only (8) was annotated with a feature (player[+3]). To be consistent, (9) should receive the same feature label. Example (10), on the other hand, is similar to (7) in that the superlative intensifies a negative evaluation (*drawback* vs. *complaint* in (7)) of a feature (*software* vs. *battery life*), however only (7) received a feature label (battery life[−3]). Given the structural and semantic similarities of the examples, one could clearly argue for adding a feature label "software[−3]" to (10).

(8) compared to everything else in this category , this is most definately [the] best [bang for the buck] . [Creative]

(9) i did a good month 's worth of research before buying this over other similar priced digital cameras , and this is [the] best [buy for the buck] . [Canon]

(10) [the] biggest [drawback that people have about the zen xtra] is the software . [Creative]

The findings of this section corroborate the claim that ISA superlatives are special indicators of product features. Their identification could simultaneously help to solve Opinion Mining tasks 1 and 2 (see above) as they frequently contain a product feature within their T or CS string, and at the same time express its associated opinion word. As this strategy for finding product features does not depend on frequency (unlike Hu&Liu's approach), ISA superlative identification also represents an efficient way of locating so-called *infrequent* features, which are also of great interest in Opinion Mining.

Automatic identification of potential product features using superlatives

Having established a positive correlation between ISA superlatives and product features, the following sections describe how instances of this superlative type can be automatically identified and how potential product feature strings can be extracted from them, using Hu and Liu's corpus of customer reviews as data set. The tool used to achieve this is SRE ('Superlative Relation Extractor'), a system implemented in Python3 which can be used to:

1) Identify superlatives in text;
2) Classify superlative instances according to the surface forms described in Table 1;
3) For superlatives classified as ISA-1, identify the spans of the target and comparison sets.

Initially, component 1) (called 'SUP-Finder') is used to find superlative instances in Hu and Liu's corpus of customer reviews. Next, the Classifier in 2)

Table3: List of abbreviations

Abbreviation	Description
CS	Comparison set of a superlative comparison
T	Target of a superlative comparison
SRE	Superlative Relation Extractor
SUP-Finder	Component of SRE used to identify superlatives in text
SUP-Classifier	Component of SRE used to classify superlative instances according to the surface forms described in Table 1
ISA1-Identifier	SUP-Classifier module used to identify ISA1-superlatives
T/CS-Identifier	Component of SRE used to identify the spans of the target and comparison sets of superlatives classified as ISA-1
CS-Identifier	Sub-component of T/CS-Identifier used to identify comparison set spans of ISA-1 superlatives
T-Identifier	Sub-component of T/CS-Identifier used to identify target spans of ISA-1 superlatives
CSDet	Determinative phrase of the superlative NP, e.g. *the* in *the best TV on the market*
CSHead	Head of the superlative NP, e.g. *TV* in *the best TV on the market*

('SUP-Classifier') is used to identify[4] ISA-1 types among the retrieved superlatives, which are then input into component 3) ('T/CSIdentifier') to extract potential product feature strings (which have been shown to occur as substrings of the target or comparison set spans). Table 3 shows an overview of common abbreviations used in the following sections.

The SRE tool was originally developed on a corpus of Wikipedia texts (TextWiki corpus; Scheible, 2008). It employs a rule-based approach based mainly on tag sequences and dependency relations (using the output of the C&C tools, cf. Clark and Curran, 2004). SRE employs rules rather than machine learning due to the relatively small size of the gold-standard data set and the low frequency of some superlative types, which would represent a problem for a learner. An additional difficulty concerns the fact that the tools used to obtai n the tags and dependency relations will have been optimised to correctly tag frequently occurring phenomena in its target text type, in order to achieve the highest possible performance score. As superlatives are relatively low frequency phenomena, with most types occurring far down the end of low frequency patterns (part of "the long tail"), even a relatively high-performance tagger like C&C may perform poorly at tagging them, because it will make little difference to the tagger's overall performance score. SRE's approach involves highly flexible and fine-tuned rules which can take these factors into account wherever necessary.

The following sections describe the three components of SRE and assess their suitability for the purpose of identifying potential product features in customer reviews. As SRE was originally developed on Wikipedia texts, its performance is expected to be affected by the non-standard nature of the data and the tagging/parsing errors that are likely to result from this.

[4] SRE is freely available upon request (email the author of this paper at `Silke.Scheible@manchester.ac.uk`.)

Superlative detection

Method

As a first step, superlatives in the corpus are automatically identified using the SUP-Finder component of SRE. As stated earlier, superlatives are derived from their base adjective/adverb in two different ways: inflectionally or analytically. In the first case, t he inflectional suffix -est is appended to the base form of the adjective or adverb (e.g. *largest*), while in the second case they are preceded by the analytical markers *most/least* (e.g. *most beautiful*). In addition, there is a (limited) number of irregular forms, such as *best*, *worst*, or *furthest*.

Previous automatic approaches to identifying superlatives have mainly focussed on techniques involving a search for the POS tags JJS and RBS (e.g. Bos and Nissim, 2006), usually without carrying out a detailed error analysis due to the large amount of manual intervention that is required for a gold standard. The SUP-Finder tool aims to improve on the POS-based approach by using a pattern matcher based on regular expressions and a list of "superlative distractors" (i.e. a list of clear cases of non-superlatives, such as *nest*, *protest*, or *honest*), which are excluded from consideration. As superlatives form a well-defined class with a limited number of irregular forms, this pattern-based search works very well, and has been shown to outperform a POS-based approach by 2-3% with 99.0% precision and 99.8% recall[5] on Wikipedia texts (Scheible, 2009).

Results and discussion

Unlike the POS-based approach, which has been optimised to work well on a particular text type, SUP-Finder is independent from text type and can be assumed to work equally well on customer review data. With its recall value nearing 100%, SUP-Finder was only assessed for precision in this study. The list of 231 superlatives returned by the tool was manually checked. Only one false positive was found, which had been missing from the list of "superlative distractors" (hobbiest, a mistyped version of hobbyist). The precision value is therefore 99.6% (230/231).

Identifying ISA superlatives

Method

The task of the second component of SRE, SUP-Classifier, is to classify superlatives as ISA-1, DEF, INDEF, etc. SUP-Classifier consists of a cascade of modules, each of which applies a set diagnostic tests to determine which class a given superlative instance belongs to. Here the focus is on the module that identifies an instance as belonging to ISA-1, called ISA1-Identifier.[6] This module requires substantial syntactic information, for example on whether the superlative form is bound in a definite NP, and if so, what the indices of the NP head and the determiner are. Furthermore, as the target of comparison needs to be explicitly mentioned in the sentence (cf. Section 3), the ISA1-Identifier component makes extensive use of the Grammatical Relations output of the C&C parser. Two main cases

[5] The only error affecting recall was due to incorrect tokenisation of quotes.

[6] Due to the low frequency of ISA-2 types, I will restrict this investigation to ISA-1 types only.

Table4: GR output for "The Panasonic is the best TV."

Row	GR output
1	`(det Panasonic_1 The_0)`
2	`(ncmod _ TV_5 best_4)`
3	`(det TV_5 the_3)`
4	`(xcomp _ is_2 TV_5)`
5	`(ncsubj is_2 Panasonic_1 _)`

are distinguished: Instances where the IS-A relation between target and comparison set is expressed via the verb "to be", or via apposition. The strategy for the former case is as follows:

- Step 1: Locate the position of the comparison set head (`CSHead`) within the sentence
- Step 2: Test whether the relation word between the `CSHead` and its dependant is a form of "to be"
- Step 3: Find the corresponding target entity

If all three steps succeed, the instance is classified as ISA-1. The first step is addressed by testing whether the head of the superlative NP (`CSHead`) occurs in subject (`ncsubj`) or complement (`xcomp`) position, as for example in (11).

- (11) The Panasonic is the best [TV]CSHead.

The output of the C&C parser for this sentence is shown in Table 4. To fulfil Step 1, the Identifier first searches for a GR tuple where CSHead (here: TV) stands in an xcomp position (Row 4 in Table 4). Step 2 is then met by checking if the item in the second slot of this tuple is a form of "to be". If it is, Step 3 is addressed by searching the GR list for another tuple where the identified verb stands in an ncsubj relation with another word (the suspected target, cf. Row 5).

Instances where the ISA relation is expressed via apposition receive a different treatment, and the following general steps are applied:

- Step 1: Test whether `CSHead` stands in a `conj` relation with any item (but excluding instances of *and*)
- Step 2: Search the GR List for the "linked" item (the suspected target)
- Step 3: Locate the position of the `CSHead` and the target in the sentence (`ncsubj` or `dobj`)

First, the list of Grammatical Relations is searched for tuples where `CSHead` stands in a `conj` relation with another word. For example, in the following sentence, Step 1 identifies the comma with index 2 as potential appositive conjunction (cf. Table 5, Row 4):

(12) The Panasonic, the best [TV]CSHead, has a PC video port.

Table5: GR output for "The Panasonic, the best TV, has a PC video port."

Row	GR output
1	(det Panasonic_1 The_0)
2	(ncmod _ TV_5 best_4)
3	(det TV_5 the_3)
4	(conj ,_2 TV_5)
5	(conj ,_2 Panasonic_1)
6	(ncmod _ port_11 video_10)
7	(ncmod _ port_11 PC_9)
8	(det port_11 a_8)
9	(dobj has_7 port_11)
10	(ncsubj has_7 TV_5 _)
11	(ncsubj has_7 Panasonic_1 _)

Table6: Results of SUP-Classifier

Class	Precision	Recall	F-measure
ISA-1	(53/56) 94.6%	(53/62) 85.5%	89.8%
Baseline	(33/115) 28.7%	(33/62) 53.2%	37.3%

Adressing Step 2, the Identifier then searches for another tuple in the list of Grammatical Relations with `conj` in first position and the comma with index 2 in second position (finding the tuple in Row 5). This is identified as a potential "linked" target. Step 3 distinguishes appositions like (12), which appear in subject position, from cases like (13), which appear in object position.

(13) I decided to order the Panasonic, the best TV.

A test is carried out to determine whether both target and `CSHead` stand in a `ncsubj` position via the same word. Rows 10 and 11 in Table 5 show that both "TV" and "Panasonic" stand in the required relation via the word "has" with index 7. The Identifier therefore concludes that the superlative appears in an apposition and classifies it as ISA-1.

Results

SUP-Classifier is tested on the output of SUP-Identifier, i.e. all superlative-containing sentences in Hu and Liu's corpus (230 altogether).[7] The results are displayed in Table 6.

The results show that SUP-Classifier clearly outperforms a random baseline system. With 94.6% precision and 85.5% recall, it can be reliably used to identify ISA-1 superlatives in customer reviews.

[7] However, five of the 230 instances were excluded from evaluation as the C&C parser failed to parse them.

Discussion

The non-standard nature of the data in customer reviews does not seem to have had the anticipated negative effect on the performance of the Classifier. Surprisingly, its performance is better on this text type than on the corpus of Wikipedia texts used in Scheible (2009), where ISA-1 achieved 82.4% precision and 84.3% recall. A closer investigation of the gold-standard ISA-1 superlatives shows that this improvement is likely to be due to a simpler syntactic structure of ISA-1 cases in customer reviews, leading to better parser performance. The C&C tool's inability to handle non-standard language mainly affected recall. For example, (12) was classified as INDEF because the system failed to identify "it 's" as erroneous variant of the possessive pronoun "its" (incorrectly tagged as personal pronoun, PRP, and 3rd person singular present tense verb, VBZ). Example (13) was not recognised by the parser because "about" is interpreted as preposition (IN) rather than as a preceding adverb (RB).

(12) i think this is itPRP 'sVBZ biggest flaw .

(13) if you do any research into digital cameras, you 'll quickly find tha t this camera is justRB aboutIN the best value out there.

Identification of potential product feature strings

Method

The third component of SRE, T/CS-Identifier, identifies potential feature-containing strings by extracting the target and comparison set strings of ISA-1 superlatives. The tool consists of two parts: a comparison set span identifier (CS-Identifier), and a target span identifier (T-Identifier). Their goal is to identify all relevant constituents of the T and CS phrases, which is a major challenge because both can have pre- and postmodifiers, the latter of which may be restrictive or nonrestrictive (Scheible, 2008). To achieve maximum accuracy, T/CS-Identifier uses a fine-grained set of rules based on the lexical annotation output of the C&C tools. This approach was chosen as the GR output by the C&C parser proved to be unreliable due to the non-standard nature of the data. Similar problems are described by Bos and Nissim (2006). The present task assumes that both target (T) and comparison set (CS) comprise a single span. The CS span is defined as consisting of a determinative phrase (CSDet) and the main CS phrase (CSMain). To identify the determinative phrase, the tool uses a purely pattern-based approach (based on POS tags). The main CS span is determined by rules which aim to identify all pre- and postmodifiers of CSHead (cf. Section 6.2). Generally, tokens occurring between the superlative form and CSHead are included as premodifiers. Postmodifiers are identified using a set of patterns which were devised to match common types of superlative postmodifiers. Target identification involves locating the target in the sentence, and identifying all restrictive pre- and postmodifiers. The following sentences are examples of superlatives for which T/CS-Identifier is able to correctly identify the target (curly brackets) and comparison set spans (square brackets), with the product feature underlined.

Table7: Performance of T/CS-Identifier (Accuracy)

Component	SRE	Baseline
CS-Identifier	62.9%	17.7%
- CSDet	98.4%	88.7%
- CSMain 64.5%	22.6%	
T-Identifier	66.1%	37.1%

(14) i think , {apex} is [the] *best* [dvd player$_{+3}$ you can get for the price] .

(15) in my opinion [the] *worst* [issue on this phone] is {the sidemounted volume control-3}.

Result

Table 7 shows the results of running T/CS-Identifier on the ISA-1 superlatives in Hu and Liu's data set. The baseline system assumes "the" as CSDet, and the first word following the superlative as the beginning of the CSMain, and the first word tagged as NN.* in that sequence as the end. The CS span is marked as correct only if both components CSDet and CSMain are exact matches with the gold standard. The baseline target identifier chooses the sequence of NP chunks closest to the superlative as target span.

Both components clearly outperform their respective baselines.

Discussion

The majority of errors in the CSMain span were caused by the tagger/parser, in cases where a restrictive "bare" relative clause starting with the pronoun "i" follows the CSHead. In (16), the parser falsely interprets "i" as the NP head because of its non-standard spelling (which caused it to be tagged as plural noun NNS instead of personal pronoun PRP). A quick test confirmed this: Running the same sentence through the tagger with "I" capitalised resulted in the correct analysis.

(16) {this} is [the] *best* [dvd player i] 've purchased .

(17) {this} is [one of the] *nicest* [phones nokia] has made .

Similarly, in (17), the token "nokia" was tagged as common noun (NN) and not recognised as a new NP chunk ('B-NP') indicating the start of a relative or subordinate clause. In both cases, the CS span breaks off incorrectly (square brackets).

While CS-Identifier performs worse on customer reviews compared to its original domain (Wikipedia texts, where it achieved 88.8%), the situation is the reverse for T-Identifier, despite the nonstandard nature of the data (66.1% vs. 58.4% in Wikipedia). This is largely due to shorter sentences and fewer appositions, which positively affect the target location methods. Furthermore, the target heads are often pronouns ("this", "it") or simple NPs such as "Apex" with no pre- or postmodifiers (30 out of 62 instances), which do not represent a problem to T Identifier. The fact that a large proportion of targets are represented by pronouns immediately raises the question of pronoun resolution. However, a first investigation of the data suggests that the great majority of the pronouns “this” and ‘it” refer to the entity under review.[8] With respect to the goal of the current investigation (i.e. identifying product features), pronouns in the target string do not represent a problem, as most product features occur in the comparison set string.

Conclusion and future work

This article established ISA-1 superlatives as special indicators of product features in Opinion Mining, which not only contain the feature strings (in most cases as part of the CS), but also the opinion word (usually the superlative itself), addressing two Opinion Mining tasks at once. Although superlatives are of relatively low frequency, the study supports previous findings that superlatives are perceived as interesting and important by people (Scheible, 2007), and Section 5 highlights their importance in customer reviews. The study further introduced SRE as a tool to reliably identify ISA-1 superlatives automatically, and to extract from them potential product feature strings. As this strategy for finding product features does not depend on frequency, it represents an efficient way of locating *infrequent* features, which are also of great interest in Opinion Mining. SRE can be used as a stand-alone system for finding product features involving ISA comparisons, or it could be incorporated as an additional component in an existing Opinion Mining system. While this is unlikely to make a considerable improvement to the overall performance score of such a system due to the relatively low frequency of superlatives, it promises to yield accurate results that are of special interest to the users.

Having automated the detection of ISA-1 superlatives and their components, the important final question is how these results can be used to arrive at the product features they are assumed to contain. As previously mentioned, the feature is a substring of either the target or the comparison set in 34 out of the 46 instances (73.9%). As the majority of them (27) occur as part of the comparison set, one strategy would be to assume that the product feature substring is the NP-chunk containing the CSHead. This simple approach would work for 25 of the 27 cases. Crucially, as most of the errors in automatically detecting the CS span were in recognising postmodification, product features can still be correctly identified as they only require identification of the CSHead chunk.

Finally, while this article has focused on the role of ISA-1 superlatives in Opinion Mining, another interesting and potentially useful class is represented by DEF, illustrated by (18) and (19), which express positive statements about the features “image quality” and “lens adapter”, respectively.

(18) overall , the g3 delivers what must be considered the *best* image quality of any current > 4 megapixel digicams , from a detail , tonal balance and color response point of view .

(19) they got the *best* lens adapter for the g3-better than canon ’s .

While the distribution of product features across the DEF class does not hint at their importance (cf. Table 2), one needs to consider that the DEF class is based

[8] This claim would however have to be verified by a thorough investigation of the context.

on surface forms and contains a variety of different semantic types, of which only the so-called "relative set comparisons" type may be of interest. Future work will therefore involve finding techniques to distinguish this type from the other semantic types found in the DEF class.

References

Bos, J., Nissim, M. (2006): An Empirical Approach to the Interpretation of Superlatives. In: *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing.* Stroudsburg (Philadelphia, USA) : Association for Computational Linguistics, pp. 9–17. (ISBN 1-932432-73-6.)

Carenini, G., Chi Kit Cheung, J. (2008): Extractive vs. NLG-based Abstractive Summarization of Evaluative Text: The Effect of Corpus Controversiality. In: *INLG-08 : International Conference on Natural Language Generation : Proceedings of the Conference.* Stroudsburg (Philadelphia, USA) : Association for Computational Linguistics, pp. 33–41.

Carenini, G., Ng, R., Pauls, A. (2006): Multi-document Summarization of Evaluative Text. In: *EACL-2006 :11th Conference of the European Chapter of the Association for Computational Linguistics : Proceedings of the Conference.* Stroudsburg (Philadelphia, USA) : Association for Computational Linguistics, pp. 305–312. (ISBN 1-932432-59-0.)

Carenini, Ng, R. T., Zwart, E. (2005): Extracting Knowledge from Evaluative Text. In: *Proceedings of the Third International Conference on Knowledge Capture : K-CAP '05.* New York : The Association for Computing Machinery, pp. 11–18. (ISBN 1-59593-163-5.)

Clark, S., Curran, J. R. (2004): Parsing the WSJ using CCG and log-linear Models. In: *ACL '04 : 42nd Annual Meeting of the Association for Computational Linguistics : Proceedings of the Conference.* Stroudsburg (Philadelphia, USA) : Association for Computational Linguistics, pp. 103–110.

Ganapathibhotla, M., Liu, B. (2008): Mining Opinions in Comparative Sentences. In: *Coling 2008 : 22nd International Conference on Computational Linguistics : Proceedings of the Conference. Volume 1.* Stroudsburg (Philadelphia, USA) : Association for Computational Linguistics, pp. 241–248. (ISBN 978-1-905593-44-6.)

Hatzivassiloglou, V., Wiebe, J. M. (2000): Effects of Adjective Orientation and Gradability on Sentence Subjectivity. In: *18nd International Conference on Computational Linguistics : Coling 2000 : Proceedings of the Conference. Vol. 1.* Saarbrücken : DFKT, pp. 299–305. (ISBN 1-55860-717-X.)

Hu, M., Liu, B. (2004): Mining Opinion Features in Customer Reviews. In: *Proceedings of the Nineteenth National Conference on Artificial Intelligence.* Palo Alto (California, USA) : Association for the Advancement of Artificial Intelligence, pp. 755–760. (ISBN 978-0-262-51183-4.)

Jindal, N., Liu, B. (2006): Mining Comparative Sentences and Relations. In: *Proceedings of the Nineteenth National Conference on Artificial Intelligence.* Palo Alto (California, USA) : Association for the Advancement of Artificial Intelligence, pp. 1331–1336. (ISBN 978-1-57735-281-5.)

Pang, B., Lee, L. (2008): Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, 2(1-2), pp. 1–135.

Popescu, A., Etzioni, O. (2005): Extracting Product Features and Opinions from Reviews. In: *HLT/EMNLP 2005 : Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing : Proceedings of the Conference.* Stroudsburg (Philadelphia, USA) : Association for Computational Linguistics, pp. 339–346.

Scheible, S. (2007): Towards a Computational Treatment of Superlatives. In: *ACL 2007 : Proceedings of the Student Research Workshop.* Stroudsburg (Philadelphia, USA) : Association for Computational Linguistics, pp. 67–72.

Scheible, S. (2008): Annotating Superlatives. In: *Proceedings of the Sixth International Language Resources and Evaluation Conference (LREC '08).* Marrakech (Morocco) : European Language Resources Association (ELRA), pp. 923–928.

Scheible, S. (2009): *A Computational Treatment of Superlatives.* (PhD thesis.) Edinburgh : University of Edinburgh, 244 pp.

Development of Japanese WordNet Affect for Analysing Emotions in Text

Yoshimitsu Torii
Precision and Intelligence Laboratory, Tokyo, Institute of Technology, 4259 Nagatsuta-cho, Midori-ku,, 226-8502 Yokohama, Japan
e-mail: `torii@lr.pi.titech.ac.jp`

Dipankar Das
Department of Computer Science and Engineering, Jadavpur University, 188, Raja S.C. Mullick Road, 700032 Kolkata, India
e-mail: `dipankar.dipnil2005@gmail.com`

Sivaji Bandyopadhyay
Department of Computer Science and Engineering, Jadavpur University, 188, Raja S.C. Mullick Road, 700032 Kolkata, India
e-mail: `sivaji_cse_ju@yahoo.com`

Manabu Okumura
Precision and Intelligence Laboratory, Tokyo, Institute of Technology, 4259 Nagatsuta-cho, Midori-ku,, 226-8502 Yokohama, Japan
e-mail: `oku@pi.titech.ac.jp`

Abstract

This paper reports on the extended task of analysing emotions in Japanese based on sense weight based scoring techniques. The previous attempt was carried out in developing Japanese *WordNet Affect* from the English *WordNet Affect* lists with the help of English *SentiWordNet* and Japanese *WordNet*. Expanding the available synsets of the English *WordNet Affect* using *SentiWordNet*, we performed the translation of the expanded lists into Japanese based on the synsetIDs in the Japanese *WordNet*. A baseline system for emotion analysis of Japanese sentences was developed based on the Japanese *WordNet Affect*. The incorporation of morphology also improved the performance of the system. Overall, the system achieved average *precision*, *recall* and *F-scores* of 32.76%, 53% and 40.49% respectively on 89 sentences of the Japanese judgment corpus and 83.52%, 49.58% and 62.22% on 1000 translated Japanese sentences of the *SemEval 2007* affect sensing test corpus. Different experimental outcomes considering different ranges of emotion scores were conducted on the *SemEval 2007* corpus. The present attempt develops the Japanese SentiWordNet with the help of English SentiWordNet and Japanese WordNet and shows that the sense weight-based scoring techniques extracted from Japanese SentiWordNet outperform the word level baseline system even including the knowledge of morphology. The first method is based on the fixed sense-tag weights that are calculated using Japanese *SentiWordNet*. Instead of using the fixed sense-tag weights, the second method calculates the lexical tag weights for each in-

[1] This paper is a slightly modified version of the paper published in the Proceedings of WASSA-2011, see `http://aclweb.org/anthology-new/W/W11/W11-1710.pdf`.

dividual word using Japanese *SentiWordNet*. The last, namely hybrid, method combines both the first and the second methods. The hybrid method considers the fixed sense-tag weights of the first method when no lexical level match is found using the second method. An averaging technique is applied to produce six sense weight scores or emotion scores of a sentence by cumulating the sense-tag weights of its word level constituents. The best emotion tag corresponding to the maximum obtained sense scores is assigned to the sentences. Finally, the hybrid method followed by the post-processing technique outperforms the other two methods by achieving an average *F-score* of 67.89% on the 1,000 translated Japanese test sentences of the *SemEval 2007* affect sensing corpus with respect to six emotions.

Key words

Japanese WordNet Affect; SentiWordNet; Emotions; SemEval 2007; Judgment Corpus

Introduction

Human-machine interface technology has been investigated for several decades. Scientists have found that emotion technology can be an important component in artificial intelligence (Salovey and Mayer, 1990). Emotions, of course, are not linguistic things. However the most convenient access that we have to them is through language (Strapparava and Valitutti, 2004). Natural language texts not only contain informative contents, but also less or more attitudinal private information including emotions. In recent times, research activities in the areas of emotion in natural language texts and other media have been gaining ground under the umbrella of subjectivity analysis (Wiebe et al., 2005) and affect computing (Strapparava and Mihalcea, 2007). The reason may be the explosive growth of the social media content on the Web in the past few years. People can now post reviews of products at merchant sites and express their emotions on almost anything in discussion forums, emails, chat, blogs, twitter and on social network sites.

The classification of reviews (Turney, 2002), newspaper articles (Lin et al., 2007), Question Answering systems (Wiebe et al., 2005) and modern Information Retrieval systems (Pang and Lee, 2008; Sood and Vasserman, 2009) have already incorporated sentiment and/or emotion analysis within their scope. The majority of subjective analysis methods that are related to opinion or emotion are based on textual keywords spotting and therefore explores the necessity of building specific lexical resources. *SentiWordNet* (Baccianella et al., 2010), a lexical resource that is used in opinion mining and sentiment analysis assigns *positive*, *negative* and *objective* scores to each synset of *WordNet* (Miller, 1995). A subjectivity wordlist (Banea et al., 2008) assigns words with the strong or weak subjectivity and prior polarities of the types *positive*, *negative* and *neutral*.

Major studies on opinion mining and sentiment analyses have been attempted with more focused perspectives rather than fine-grained emotions (Quan and Ren, 2009). The extraction and annotation of subjective terms started with machine learning approaches (Hatzivassiloglou and McKeown, 1997). Some well-known sentiment lexicons have been developed, such as the subjective adjective list (Baroni and Vegnaduzzo, 2004), English *SentiWordNet* (Esuli et al., 2006), Taboda's adjective list (Voll and Taboda, 2007), Subjectivity Word List (Banea et al., 2008), etc. The affective lexicon (Strapparava and Valitutti, 2004), one of the most efficient resources of emotion analysis, contains words that convey emotion. It is a small well-used lexical resource but valuable for its affective annotation.

Andreevskaia and Bergler (2006) present a method for extracting *positive* or *negative* sentiment-bearing adjectives from *WordNet* using the Sentiment Tag Extraction Program (STEP). The methods proposed in Wiebe and Riloff, 2006, automatically generate resources for subjectivity analysis for a new target language from the available resources for English. Two techniques have been proposed for the generation of a target language lexicon from the English subjectivity lexicon. The first technique uses a bilingual dictionary, while the second method is a parallel corpus-based approach using existing subjectivity analysis tools for English. In contrast, instead of using any dictionary or parallel corpus, we have used the Japanese *WordNet* (Bond et al., 2009) to accomplish the translation purpose. The methods proposed in Mohammad et al. (2008), help to measure the relative sentiment score of a word and its antonym. On the other hand, an automatically generated and scored sentiment lexicon, *SentiFul* (Neviarouskaya et al., 2009), its expansion, morphological modifications and distinguishing sentiment features (propagating, reversing, intensifying, and weakening) also shows contributory results.

To the best of our knowledge all of the above mentioned resources are in English and have been used in coarse-grained sentiment analysis (e.g., *positive*, *negative* or *neutral*). The proposed method in Takamura et al. (2005) extracts semantic orientations from a small number of seed words with high accuracy in the experiments on English as well as Japanese lexicons. However it was also aimed at sentiment-bearing words. There is always a demand for automatic text analysis tools and linguistic resources for languages other than English. A recent study shows that non-native English speakers support the growing use of the Internet[2]. Instead of English *WordNet Affect* (Strapparava and Valitutti, 2004), there are a few attempts in other languages such as Russian and Romanian (Bobicev et al., 2010), Bengali (Das and Bandyopadhyay, 2010), etc. Our previous and current approaches are similar to some of these approaches but in contrast, we evaluated our Japanese *WordNet Affect* on the *SemEval 2007* affect-sensing corpus translated into Japanese. The above mentioned approaches use a bilingual dictionary whereas we have used the Japanese *WordNet* for translation. Translation based on Japanese *WordNet* is more reliable than translation using a bilingual dictionary.

In recent trends, the application of Mechanical Turk for generating a emotion lexicon (Mohammad and Turney, 2010) shows a promising avenue of research. To avoid any monetary investment in developing an emotion lexicon, we have incorporated open source, available and accessible resources to achieve our goals.

In our previous attempt, we prepared a Japanese *WordNet Affect* from the already available English *WordNet Affect* (Strapparava and Valitutti, 2004). Entries in the English *WordNet Affect* are annotated using Ekman's (1993) six emotional categories (*joy, fear, anger, sadness, disgust and surprise*). The collection of the English *WordNet Af-*

[2] `http://www.internetworldstats.com/stats.htm`

fect[3] synsets that are used in the present work was provided as a resource in the '*Affective Text*' shared task of the *SemEval-2007* Workshop. The shared task focused on text annotation with affective tags (Strapparava and Mihalcea, 2007). We have not considered the problems of lexical affect representation or discussed the differences between emotions, cognitive states and affects in developing Japanese *WordNet Affect*.

The six *WordNet Affect* lists that were provided in the shared task contain only 612 synsets in total with 1,536 words. The words in each of the six emotion lists have been observed to be not more than 37.2% of the words present in the corresponding *SentiWordNet* synsets. Hence, these six lists were expanded with the synsets retrieved from the English *SentiWordNet* (Baccianella et al., 2010) to have an adequate number of emotion-related word entries. We assumed that the new sentiment-bearing words in English *SentiWordNet* might have some emotional connotation in Japanese. However, the part-of-speech (POS) information for each of the synsets was kept unchanged during expansion of the lists. The numbers of entries in the expanded word lists were increased by 69.77% and 74.60% at synset and word levels, respectively. We mapped the synsetID of the *WordNet Affect* lists onto the synsetID of the WordNet 3.0[4]. This mapping helps in expanding the *WordNet Affect* lists with the recent version of *SentiWordNet 3.0*[5] as well as translating with the Japanese *WordNet* (Bond et al., 2009).

Japanese *WordNet*[6], a freely available lexical resource, is being developed based on the English *WordNet*. The synsets of the expanded lists were automatically translated into Japanese equivalent synsets based on the synsetID. Some synsets (e.g., 00115193-a *huffy, mad, sore*) were not translated into Japanese as there were no equivalent synset entries in Japanese *WordNet* for those affect synsets.

Primarily, we developed a baseline system based on the Japanese *WordNet Affect* and carried out the evaluation on a Japanese judgement corpus of 89 sentences. The system achieved an average *F-score* of 36.39% with respect to six emotion classes. We also incorporated a morphological knowledge of the emotion words into the baseline system using an open source Japanese morphological analyser[7]. The performance of the system was increased by 4.1% in average *F-score* with respect to six emotion classes.

The lack of an emotion corpus in Japanese motivated us to apply an open source Google translator[8] to build a Japanese emotion corpus from the available emotion corpus in English. The English *SemEval-2007* affect-sensing corpus contains the trial and test sets of 250 and 1,000 sentences of news headlines. Each sentence of the corpus is annotated with six emotion scores for Ekman's six emotion types and three valence scores for *positive, negative* or *neutral* types. In the previous task, we considered that each sentence is to be assigned with a single sentential emotion tag based on the maximum emotion score out of six annotated emotion scores. The baseline system based on the Japanese *WordNet Affect* achieved the aver-

[3] `http://www.cse.unt.edu/~rada/affectivetext/`

[4] `http://wordnet.princeton.edu/wordnet/download/`

[5] `http://sentiwordnet.isti.cnr.it/`

[6] `http://nlpwww.nict.go.jp/wn-ja/index.en.html`

[7] `http://mecab.sourceforge.net/`

[8] `http://translate.google.com/#`

age *precision*, *recall* and *F-score* of 83.52%, 49.58% and 62.22%, respectively on 1,000 translated test corpus. It has to be mentioned that the inclusion of morphological processing improved the performance of the system. Different experiments were carried out by selecting different ranges of annotated emotion scores. It was observed that selecting lower emotion scores, the number of sentential instances in each of the six emotion categories was increasing and the performance of the system was showing subsequent improvement. Error analysis suggested that though the system performed satisfactorily in identifying the sentential emotions based on the available words of the Japanese *WordNet Affect*, the system suffered from the translated version of the corpus. In addition to that, the Japanese *WordNet Affect* also needs improvement in terms of coverage.

In our present extended task, the Japanese SentiWordNet is being developed by mapping the synsets of English SentiWordNet with Japanese WordNet via English WordNet. We have calculated the polarised sense weights of six emotion tags using Japanese *SentiWordNet*. Three different sense weight-based scoring techniques have been employed for assigning six emotion scores to each of the sentences based on their word level emotion-tagged constituents. The first method is based on the fixed sense-tag weights that are calculated using the Japanese *SentiWordNet*. Instead of depending on the fixed sense-tag weights, the second method calculates the tag weights of each individual word by directly searching them in the Japanese *SentiWordNet*. In contrast to these methods, the last method mimics the second method but the only difference is that the words that are absent in the Japanese *SentiWordNet* consider the fixed sense-tag weights of the first method. All of the methods assume the sense-tag weight of the *neutral* tag is zero. An averaging technique is applied to produce six sense weight scores or emotion scores of a sentence from the sense-tag weights of its word level constituents. Only one sentential emotion tag is assigned to each of the sentences based on the maximum emotion score obtained by the system. The evaluation of assigning emotion tags to the sentences achieves an average F-score of 64.71% on the development set of the *SemEval 2007* corpus. The post-processing technique has been applied on the development set for handling negation words and the F-score improved to 66.14%. The evaluation on 250 test sentences yields an overall F-score of 67.89%.

The rest of the paper is organised as follows. Different developmental phases of the Japanese *WordNet Affect* and Japanese SentiWordNet are described in Section 2. Preparation of the translated Japanese corpus, different morphology-based experiments on the Japanese judgment corpus and the translated corpus, experiments based on the annotated emotion scores of the translated corpus and subsequent evaluations are elaborated in Section 3. Section 4 discusses the sense-based scoring techniques for identifying sentence-level emotion tags. Finally Section 5 concludes the paper.

Development Phases

WordNet Affect

The English *WordNet Affect* (Strapparava and Valitutti, 2004), based on Ekman's (1993) six emotion types (*joy, fear, anger, sadness, disgust, surprise*) is a small lexical re-

Figure 1: Linking between the synsets of *WordNet Affect* and *WordNet*

WordNet Affect:
n#05587878 anger choler ire
a#02336957 annoyed harassed harried pestered vexed
WordNet:
07516354-n anger, ire, choler
02455845-a annoyed harassed harried pestered vexed
Linked *Synset ID* with *Affect ID*:
n#05587878 ↔ 07516354-n anger choler ire
a#02336957 ↔ 02455845-a annoyed harassed harried pestered vexed

source compared to the complete *WordNet* (Miller, 1995) but its affective annotation helps in emotion analysis. Some collection of *WordNet Affect* synsets was provided as a resource for the shared task of *Affective Text* in *SemEval-2007* (Strapparava and Mihalcea, 2007). The whole data is provided in six files named for the six emotions. Each file contains a list of synsets and one synset per line. An example synset entry from *WordNet Affect* is shown as follows.

a#00117872 angered enraged furious infuriated maddened

The first letter of each line indicates the part of speech (POS) and is followed by the *affectID*. The representation was simple and easy for further processing. We retrieved and linked the compatible synsetID from the recent version of *WordNet 3.0* with the *affectID* of the *WordNet Affect* synsets using an open source tool[9]. The linking of two *WordNet Affect* synsets with their corresponding synsets of *WordNet 3.0* is shown in Figure 1. The differences between emotions, cognitive states and affects were not analysed during translation. Our main focus in the task was to develop an equivalent resource in Japanese for analysing emotions.

Expansion of WordNet Affect using SentiWordNet

It was observed that the *WordNet Affect* (Strapparava and Valitutti, 2004) contains fewer emotion word entries. The six lists provided in the *SemEval 2007* shared task contain only 612 synsets in total with 1,536 words. The detailed distribution of the emotion words as well as the synsets in six different lists according to their POS are shown in Table 1.

Hence, we expanded the lists with adequate number of emotion words using *SentiWordNet* (Baccianella et al., 2010) before attempting any translation of the lists into Japanese. *SentiWordNet* assigns each synset of *WordNet* with two coarse-grained subjective scores such as *positive*, *negative* along with an *objective* score. *SentiWordNet* contains more number of coarse-grained emotional words than *WordNet Affect*. We assumed that the translation of the coarse-grained emotional words into Japanese might contain more or less fine-grained emotion words. One example entry of the *SentiWordNet* is shown below. The POS of the entry is followed by a *synset ID*, *positive* and *negative* scores and synsets containing sentiment words.

[9] http://nlp.lsi.upc.edu/web/index.php?option=com_content&task=view&id=21&Itemid=59

Figure 2: Expansion of *WordNet Affect* synset using *SentiWordNet*

Linked*Affect word:*
n#05587878↔ *07516354-nanger choler ire*newline *SentiWordNet* synsets that include the word "*anger*":
07516354-n anger, ire, choler
14036539-n angriness, anger
00758972-n anger, ira, ire, wrath
01785971-v anger
01787106-v see_red, anger
SentiWordNet synsets that include the word "choler":
07552729-n fretfulness, fussiness, crossness, petulance, peevishness, irritability, choler
05406958-n choler, yellow_bile
Expanded*Affect word*:
n#05587878 ↔ *07516354-n anger choler ire 14036539-n angriness anger 00758972-n anger ira, ire wrath 01785971-vanger ... 05406958-n choler*

SentiWordNet: a 121184 0.25 0.25
infuriated#a#1 furious#a#2 maddened#a#1 enraged#a#1 angered#a#1

Our aim was to increase the number of emotion words in the *WordNet Affect* using *SentiWordNet*. Both of the two resources were developed from the *WordNet* (Miller, 1995). Hence, each word of the *WordNet Affect* is replaced by the equivalent synsets retrieved from *SentiWordNet* if the synset contains that emotion word. The POS information in the *WordNet Affect* is kept unchanged during expansion. For example, in Figure 2, the word '*anger*' in synset '*07516354-n*' is linked with the synsets '*14036539-n*' "*angriness*, *anger*", '00758972-n' "*anger*, *ira*, *ire*, *wrath*", '*01785971-v*' "*anger*", '*01787106-v*' "*see_red*, *anger*", etc., and therefore the linked words and synsets are appended to the existing word '*anger*'. The distributions of expanded synsets and words for each of the six emotion classes based on four different POS types (*noun N*, *verb V*, *adjective Adj.* and *adverb Adv.*) are shown in Table 1 and Table 2. However we have kept the duplicate entries at synset level for identifying the emotion-related scores in our future attempts by utilising the already associated *positive* and *negative* scores of *SentiWordNet*. The percentage of entries in the updated word lists was increased by 69.77 and 74.60 at synset and word levels, respectively.

In case of word ambiguity during the replacement of the words in WordNet affect synsets, some spurious senses appeared in some synsets that represent a non-appropriate meaning. However, it was observed that in the case of emotion words, this phenomenon is not frequent because the direct emotion words are not very ambiguous.

Translation of Expanded WordNet Affect into Japanese

We mapped the *affectID* of the *WordNet Affect* to the corresponding *synsetID* of the *WordNet 3.0*. This mapping helps to expand the *WordNet Affect* with the recent version of *SentiWordNet* 3.0 as well as translating the expanded lists into Japanese using the Japanese *WordNet* (Bond et al., 2009).

Table 1: Number of POS-based Synset entries in six *WordNet Affect* lists before and after updating using *SentiWordNet*

	WordNet Affect List Synset Entries [After updating using SentiWordNet]			
Emotion Classes	Noun	Verb	Adjective	Adverb
anger	48 [*198*]	19 [*103*]	39 [*89*]	21 [*23*]
disgust	3 [*17*]	6 [*21*]	6 [*38*]	4 [*5*]
fear	23 [*89*]	15 [*48*]	29 [*62*]	15 [*21*]
joy	73 [*375*]	40 [*252*]	84 [*194*]	30 [*45*]
sadness	32 [*115*]	10 [*43*]	55 [*129*]	26 [*26*]
surprise	5 [*31*]	7 [*42*]	12 [*33*]	4 [*6*]

Table 2: Number of POS-based Word entries in six *WordNet Affect* lists before and after updating using *SentiWordNet*

	WordNet Affect List Word Entries [After updating using SentiWordNet]			
Emotion Classes	Noun	Verb	Adjective	Adverb
anger	99 [*403*]	64 [*399*]	120 [*328*]	35 [*50*]
disgust	6 [*21*]	22 [*62*]	34 [*230*]	10 [*19*]
fear	45 [*224*]	40 [*243*]	97 [*261*]	26 [*49*]
joy	149 [*761*]	122 [*727*]	203 [*616*]	65 [*133*]
sadness	64 [*180*]	33 [*92*]	169 [*779*]	43 [*47*]
surprise	8 [*28*]	28 [*205*]	41 [*164*]	13 [*28*]

As the Japanese *WordNet*[10] is freely available and it is being developed based on the English *WordNet*, the synsets of the expanded lists were automatically translated into Japanese equivalent synsets based on the *synsetIDs*. The number of translated Japanese words and synsets for the six affect lists are shown in Table 3 and Table 4, respectively. The following are some translated samples that contain word level as well as phrase level translations.

07510348-n surprise rightarrow 愕き, 驚き

*07503260-n disgust rightarrow*むかつき, 嫌悪

*07532440-n unhappiness, sadness rightarrow*不仕合せさ, 哀情, 悲しみ, 不幸せさ, 不幸さ, 不幸せ, 不仕合わせ, 哀しみ, 不仕合せ, 不幸, 悲しさ, 不仕合わせさ, 哀しさ

07527352-n joy, joyousness, joyfulness → ジョイ, 愉楽, うれしいこと, 慶び, うれしさ, 歓び, 悦楽, 歓, 嬉しさ, 欣び, 楽しいこと, 喜び, 楽しさ, 悦び, 愉悦

Translation of SentiWordNet into Japanese

In the present extended task, we have prepared the Japanese SentiWordNet using the English SentiWordNet and the Japanese WordNet. The English *SentiWordNet* (Baccianella et al., 2010), an important resource in opinion mining and sentiment analysis assigns three sentiment scores such as *positive*, *negative* and *objective* to each synset of *WordNet*. As the Japanese WordNet is also aligned with the English

[10] `http://nlpwww.nict.go.jp/wn-ja/index.en.html`

Table 3: Number of POS-based translated word entries in six Japanese *WordNet Affect* lists

Emotion Classes	Translated *WordNet Affect* list and SentiWordNet in Japanese (#Words)			
	Noun	Verb	Adjective	Adverb
anger	861	501	231	9
disgust	49	63	219	10
fear	375	235	334	104
joy	1959	1831	772	154
sadness	533	307	575	39
surprise	144	218	204	153
SentiWordNet	2856	346	12,102	233
SentiWordNet (pos/neg)	826	167	5,423	104

WordNet at synset level, we mapped the English synsets of SentiWordNet onto the Japanese WordNet using the intermediate synsetIDs of English WordNet. It was observed that the total number of non-translated synsets was significantly higher in comparison with the total number of translated synsets. The numbers of POS-based translated synsets, words and phrases are shown in Table 3 and Table 4. However, a crucial fact was found with respect to the subjective (i.e., *positive* and/or *negative*) and objective scores of the English *SentiWordNet* synsets. Only 17,996 synsets that contain *positive* and/or *negative* scores are present in the English *SentiWordNet* lexicon and out of these, 32.33% of synsets have been translated in Japanese. The overall translation yields 55.35% of the synsets that contain scores for either *positive* or *negative* or both types of sentiments. The reason for non-translated synsets may be that the Japanese WordNet is being developed and not yet completed with respect to the English WordNet.

Analysing Translation Errors

Some *SentiWordNet* synsets (e.g., 00115193-a *huffy, mad, sore*) were not translated into Japanese as there are no equivalent synset entries in the Japanese *WordNet*. There were a large number of word combinations, collocations and idioms in the Japanese *WordNet Affect*. These parts of synsets show problems during translation and therefore manual translation was carried out for these types. There are some of the English synsets that were not translated into Japanese. For example, the synset '07517292-n *lividity*' contains only one English word that was not translated into Japanese. However an equivalent gloss of the word '*lividity*' that is present in the Japanese *WordNet* is "*a state of fury so great the face becomes discoloured*". One of the reasons for such translation problems may be that no equivalent Japanese word sense is available for such English words.

WordNet Affect-based Evaluation

Knowledge resources can be leveraged in identifying emotion-related words in text and the lexical coverage of these resources may be limited given the informal nature of online discourse (Aman and Szpakowicz, 2007). In general, the identification of

Table 4: Number of *translated* and *non-translated* synset entries and *morphemes* including *words* and *phrases* in six Japanese *WordNet Affect* lists

	Japanese *WordNet Affect* list and SentiWordNet			
	Translated	Non-Translated	Translated	Translated
Emotion Classes	(#Synset)	(#Synset)	(#Word)	(#Phrase)
anger	254	159	1033	450
disgust	57	24	218	97
fear	146	74	615	315
joy	628	238	2940	1273
sadness	216	97	846	519
surprise	112	25	456	216
SentiWordNet	10,513	1,07,146	15,537	3,107
SentiWordNet (pos/neg)	5,819	11,877	6,520	707

the direct emotion words incorporates the lexicon lookup approach. Hence, we evaluated the developed Japanese *WordNet Affect* on a small emotional judgment corpus and *SemEval 2007* affect-sensing corpus in Japanese.

Evaluation on Japanese Judgment Corpus

The judgment corpus that is being developed by the Japan System Applications Co. Ltd.[11] contains only 100 sentences of emotional judgments. However this corpus is not an open source as yet. We evaluated our Japanese *WordNet Affect*-based baseline system on these 100 sentences and the results for each of the six emotion classes are shown in Table 5. We also incorporated the morphological knowledge in our baseline system using an open source Japanese morphological analyser[12].

The algorithm is such that if a word in a sentence is present in any of the Japanese *WordNet Affect* lists, the sentence is tagged with the emotion label corresponding to that affect list. If any word is not found in any of the six lists, each word of the sentence is passed through the morphological process to identify its root form and the root form is searched for through the Japanese *WordNet Affect* lists again. If the root form is found in any of the six Japanese *WordNet Affect* lists, the sentence is tagged accordingly. Otherwise, the sentence is tagged as non-emotional or *neutral*.

It was observed that the average *F-Score* of the baseline system improved by 4.1% with respect to the six emotion classes. Due to the lower number of sentential instances in some emotion classes (e.g., *joy*, *sadness*, *surprise*), the performance of the system gives poor results even after including the morphological knowledge. One of the reasons may be the fewer word and synset entries in some *WordNet Affect* lists (e.g., *fear*). Another reason was the lower number of sentential instances in some emotion class (e.g., *sadness*). Hence, we aimed to translate the English *SemEval 2007* affect-sensing corpus into Japanese and evaluate our system on the translated corpus.

[11] `http://www.jsa.co.jp/`

[12] `http://mecab.sourceforge.net/`

Table 5: Precision, Recall and F-Scores (in %) of the Japanese *WordNet Affect*-based system per emotion class on the Judgment corpus before and after the inclusion of morphology.

	Judgment Corpus (in %)		
	Before Morphology [*After Morphology*]		
Emotion Classes (#Sentences)	Precision	Recall	F-Score
anger (#32)	51.61 [*64.29*]	50.00 [*68.12*]	50.79 [*66.14*]
disgust (#18)	25.00 [*45.00*]	5.56 [*10.56*]	9.09 [*17.10*]
fear (#33)	NULL	NULL	NULL
joy (#3)	3.45 [*8.08*]	66.67 [*100.00*]	6.56 [*14.95*]
sadness (#5)	NULL	NULL	NULL
surprise (#9)	6.90 [*13.69*]	22.22 [*33.33*]	10.53 [*19.41*]

Evaluation on Translated SemEval 2007 Affect-Sensing Corpus

The English *SemEval 2007* affect-sensing corpus (Strapparava and Mihalcea, 2007) consists of news headlines only. Each of the news headlines is tagged with a valence score and scores for all six of Ekman's (1993) emotions. The six emotion scores for each sentence are in the range of 0 to 100.

We used the Google translator API[13] to translate the 250 and 1,000 sentences of the trial and test sets of the *SemEval 2007* corpus respectively. The experiments regarding morphology and emotion scores were conducted on the trial corpus. The final evaluation that was carried out on 1000 sentences of the test corpus produces the results shown in Table 6. The evaluation of our system is similar to the coarse-grained evaluation methodology of the *SemEval 2007* shared task on affective text. Though the evaluation was conducted for Japanese, the performance of the system improved significantly. In addition to the coarse-grained evaluation, we also carried out different experiments by selecting different ranges of emotion scores. The corresponding experimental results are also shown in Table 6. Incorporation of morphology improves the performance of the system.

On the other hand, it was observed that the performance of the system decreases by increasing the range of Emotion Scores (ES). The reason may be that the numerical distribution of the sentential instances in each of the emotion classes decreases as the range in emotion scores increases. This, in turn, decreases the performance of the system.

Japanese affect lists include words as well as phrases. We deal with phrases using a Japanese morphology tool to find affect words in a sentence and substitute an affect word into its original conjugated form. One of the main reasons for using a morphology tool was to analyse the conjugated form and to identify the phrases. For example, the Japanese word for the equivalent English word '*anger*' is '怒る (*o ko ru*)', but there are other conjugated word forms such as '怒った (*o ko tta*)' that means '*angered*' and it is used in the past tense. Similarly, another conjugated form '怒っていた (*o ko tte i ta*)' denotes the past participle form '*have angered*' of the original word '*anger*'. The morphological form of its passive sense is '怒られる (*o ko ra re ru*)' that means '*be angered*'. In addition to that, we identified the words into their original

[13] http://translate.google.com/\#

forms from their corresponding phrases by using the morpheme information. For example, the phrase '怒られる (*o ko ra re ru*)' consists of two words, one is '怒ら (*o ko ra*)', which is in an imperfective form, and the other is 'れる (*re ru*)', which is in an original form. The original form of the imperfective word '怒ら (*o ko ra*)' is '怒る (*o ko ru*)'.

It was found that some of the English multi-word phrases have no equivalent Japanese phrase available. Only the equivalent Japanese words were found in the Japanese *WordNet*. For example, the following synset contains a multi-word phrase '*see-red*'. But, instead of any equivalent phrases, only words are found in Japanese *WordNet*.

01787106-v anger, see -red $rightarrow$怒る, 憤る, 立腹

Sense Weight Score-based Evaluation

The present task also incorporates the sense weight score-based evaluation of the system. For this purpose, we have used the *positive* and/or *negative* scores of the words that are present in the synsets of Japanese *SentiWordNet*. Three different methods based on the sense weights have been considered for assigning emotion scores to the sentences. The methods consider the average weighting technique to identify the emotion tags for a sentence. Each of the sentences is assigned with a final emotion tag based on the maximum sense weight scores that are assigned by the system. Similarly, each of the sentences is already annotated with a sentence level gold standard emotion tag in the *SemEval 2007* corpus based on the maximum emotion scores that were assigned by the annotators. The system assigned emotion tags are evaluated with respect to the gold standard emotion tags and the results have shown satisfactory performance in coarse-grained evaluation.

Fixed Sense Weight-based Scoring (FSWS)

In the first method, we have chosen the basic six words in Japanese '悲しい' (*happy*), '幸せ' (*sad*), '怒り' (*anger*), '嫌悪' (*disgust*), '恐怖' (*fear*) and '驚き' (*surprise*) as the seed words corresponding to each type of emotion tag. The *positive* and *negative* scores for each synset in which each of these seed words appear are retrieved from the Japanese *SentiWordNet* and the average of the scores is fixed as the *Sense_Tag_Weight* (*STW*) of that particular emotion tag. The present sense weight-based scoring technique is based on the hypothesis that was considered in (Das and Bandyopadhyay, 2009). Table 7 shows the values of *STW* for six emotion tags. The *neutral* tag is assigned *zero* value as it does not carry any emotional sense. These sense-based tag weights (*STW*) are fixed value in nature. Each sentence is assigned with six sentence level sense weights (*SWS*) with respect to six emotions. Each of the weights is calculated by dividing the total *Sense_Tag_Weight* (*STW*) of an emotion type by the total *Sense_Tag_Weight* (*STW*) of all types of emotion present in that sentence. The sentence is assigned with a single emotion tag for which the sentence level sense weight score (*SWS*) is maximum.

$$SWS_i = (STW_i * N_i)/(\sum_{j=1}^{7} STW_j * N_j)|i \in j,$$

Table 6: Precision, Recall and F-Scores (in %) of the Japanese *WordNetAffect*-based system per emotion class on the translated Japanese *SemEval 2007* affect-sensing test corpus before and after the inclusion of morphology on different ranges of Emotion Scores (ES).

Emotion Classes	Japanese Translated SemEval 2007 Test Corpus		
(#Sentences)	Before Morphology [*After Morphology*]		
	Precision	Recall	F-Score
Emotion Score (ES) ≥0			
anger	61.01 [*68.75*]	18.83 [*31.16*]	28.78 [*42.88*]
disgust	79.55 [*85.05*]	8.35 [*16.06*]	15.12 [*27.01*]
fear	93.42 [*95.45*]	10.26 [*16.77*]	18.49 [*28.52*]
joy	69.07 [*72.68*]	57.03 [*80.30*]	62.48 [*76.29*]
sadness	83.33 [*84.29*]	10.58 [*19.54*]	18.77 [*31.67*]
surprise	94.94 [*94.94*]	7.84 [*13.65*]	14.48 [*23.99*]
Emotion Score (ES) ≥ 10			
anger	44.65 [*52.08*]	25.54 [*33.32*]	32.49 [*40.35*]
disgust	40.91 [*41.46*]	9.89 [*18.07*]	15.93 [*24.97*]
fear	77.63 [*81.82*]	13.32 [*21.42*]	22.74 [*34.03*]
joy	53.89 [*55.61*]	56.50 [*96.22*]	55.17 [*70.40*]
sadness	67.78 [*69.87*]	11.78 [*19.88*]	20.07 [*30.86*]
surprise	72.15 [*74.58*]	8.25 [*15.87*]	14.81 [*26.30*]
Emotion Score (ES) ≥ 30			
anger	21.38 [*28.12*]	39.08 [*62.45*]	27.64 [*38.59*]
disgust	2.27 [*5.04*]	3.70 [*6.72*]	2.82 [*6.15*]
fear	44.74 [*56.82*]	16.67 [*28.76*]	24.29 [*38.45*]
joy	31.48 [*33.42*]	56.86 [*97.08*]	40.52 [*50.53*]
sadness	37.78 [*69.86*]	15.60 [25.31]	22.08 [*37.22*]
surprise	17.72 [*20.34*]	8.14 [*18.56*]	11.16 [*20.35*]
Emotion Score (ES) ≥ 50			
anger	6.92 [*10.42*]	57.89 [78.02]	12.36 [*18.26*]
disgust	NIL	NIL	NIL
fear	21.05 [*29.55*]	17.98 [*31.26*]	19.39 [*30.79*]
joy	12.04 [*24.98*]	61.32 [*87.66*]	20.12 [*39.10*]
sadness	13.33 [*23.07*]	12.12 [*22.57*]	12.70 [*18.71*]
surprise	3.80 [*8.50*]	7.50 [*12.50*]	5.04 [*10.11*]

where pojemSWS$_i$ is the sentence level sense weight score for the emotion tag i and N_i is the number of occurrences of the emotion tag i in the sentence. STW_j is the *Sense_Tag_Weight* for each emotion tag j including the emotion tag i. The emotion tag corresponding to the maximum sense weight score (*SWS*) is assigned to a sentence as the probable emotion tag. It has to be mentioned that only the magnitude, not the *polarity* (*positive/negative*) that is also attached with *STW* was considered in case of calculating *SWS*.

Table 7: Six Sense-tag Weights (STWs) for six emotion tags and neutral tags.

Emotion Classes	*Sense-tag Weights* (*STW*)
anger	0.0125
disgust	(−) 0.1022
fear	(−) 0.5
joy	(−) 0.075
sadness	0.0131
surprise	0.0625
neutral	0.0

Lexical Sense Weight-based Scoring (LSWS)

In the second method, we have considered the emotion-tagged words instead of their fixed assigned sense-tag weights (as mentioned in the first method). Each emotion-tagged word is searched in the Japanese *SentiWordNet*. The *positive* and *negative* scores of the word are retrieved from the Japanese *SentiWordNet* and the average of the retrieved scores has been fixed as the *Sense_Tag_Weight* (*STW*) for that emotion word. Morphological processing of the words has also been included into the search process. If the word as well as its stem form is not found in the Japanese *SentiWordNet*, the default value is assumed as zero. In this method, the total STW_i for each emotion tag i is calculated by summing up the *STW*s of all assigned emotion tags with type i.

Hybrid Sense Weight-based Scoring (HSWS)

The third method is similar with the second method. The main difference is that this method uses the fixed sense-tag weights instead of assuming zero values for the emotion words that are not present in Japanese *SentiWordNet* in its original as well as stem forms. If an emotion-tagged word is not found in the Japanese *SentiWordNet*, the default sense-tag weight that was used in the first method is assigned for that emotion tag.

The evaluation of these three methods has been carried out on the development and test sets. The results are shown in Table 8. It has been observed that the hybrid method significantly outperforms the other two methods. The hybrid method incorporates the knowledge of individual sense weight for an emotion-tagged word as well as using the default weight for the words that have no clue in the Japanese SentiWordNet lexicon. As the hybrid method shows better performance than the other two methods, it has been applied in identifying the sentence level emotion tags.

Pre-Processing for Handling Negations

The presence of negations and their number of occurrences are both significant in assigning the final emotion tag to a sentence. The consecutive occurrence of negation words does not reverse the assigned emotion type whereas the presence of a single negation may completely change the actual emotion. For example, the

Table 8: F-score (in %) of three sense-weight based scoring methods for six emotion classes.

Emotion Classes	Japanese Translated SemEval 2007 Test Corpus		
	F-Score (in %) [Before Pre-processing for Negation Words]		
	FSWS	*LSWS*	*HSWS*
anger	56.78 [59.23]	58.22 [60.31]	65.12 [68.45]
disgust	52.09 [54.44]	54.76 [57.22]	61.07 [64.89]
fear	57.34 [59.02]	59.44 [62.10]	66.07 [69.07]
joy	53.21 [56.09]	59.38 [61.55]	63.10 [65.78]
sadness	57.02 [59.11]	60.21 [63.14]	67.36 [68.55]
surprise	58.53 [60.57]	61.37 [63.03]	66.28 [67.13]
Average	55.82 [58.07]	58.89 [61.22]	64.83 [67.31]

following sentence was tagged as "*sad*" by the system but in the gold standard *SemEval 2007* corpus, the maximum emotion score is given for "*happy*".

パリジャーナル：スモーキングなし長いトレスフランスのシック

Paris Journal: Smoking No Longer Tres Chic in France

Thus, considering the immediate presence and single occurrence of the negation word ない (*No*), the emotion tag of the sentence is reversed to "*happy*". It has to be mentioned that, the negations only play the roles in the case of two emotions such as "*happy*" and "*sad*". However in the case of other emotions, the single negation word has no role to play.

In the following sentence, two consecutive occurrences of negation words (ない (*No*) and ない (*Not*) do not change the actual emotion expressed by the sentence.

スナックで誘惑？いいえ、あなた

Seduced by Snacks? No, Not You

In this case, the system assigns the "*fear*" tag that is also the probable maximum scored emotion tag in the gold standard annotated corpus. Application of these rule-based post-processing strategies improved the F-score of the system for identifying sentence-level emotion tags. The results are shown in Table 8. Overall, the 2% 3% F-score has been improved by employing the post-processing techniques for handling negations.

Conclusion

The present paper describes the extended task of preparation of Japanese *WordNet Affect* and its evaluation on the Japanese Judgment corpus and SemEval 2007 affect-sensing corpus. The automatic approach to expanding, translating and sense disambiguation tasks reduces the manual effort. The resource is still being updated with more number of emotional words to increase the coverage. In addition to Japanese *WordNet Affect*, the Japanese SentiWordNet is also being developed and its sense-based scores have been used to identify sentential emotion tags. Our future task is to integrate more resources so that the number of emotion word entries in the Japanese SentiWordNet can be increased.

References

Aman S., Szpakowicz, S. (2007): Identifying Expressions of Emotion in Text. *TSD 2007, LNAI*, Vol. 4629, pp. 196–205.

Andreevskaia, A., Bergler, S. (2007): CLaC and CLaC-NB: Knowledge-based and corpus-based approaches to sentiment tagging. In: *4th International Workshop on Semantic Evaluations (SemEval).* Stroudsburg (Philadelphia, USA) : Association for Computational Linguistics, pp. 117–120.

Baccianella, S., Esuli, A., Sebastiani, F. (2010): SentiWordNet 3.0: An Enhanced Lexical Re-source for Sentiment Analysis and Opinion Mining. In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10).* Valletta (Malta) : European Language Resources Association (ELRA), pp. 2200–2204. (ISBN 2-9517408-6-7.)

Banea, C., Mihalcea, R., Wiebe, J. (2008): A Bootstrapping Method for Building Subjectivity Lexicons for Languages with Scarce Resources. In: *The proceedings of the Sixth International Conference on Language Resources and Evaluation 2008 [CD-ROM].* Marrakech (Morocco) : European Language Resources Association (ELRA), pp. 2764–2767.

Baroni, M., Vegnaduzzo, S. (2004): Identifying Subjective Adjectives Through Web-based Mutual Information. In: Buchberger E. (ed.) *Proceedings of the German Conference on NLP (KONVENS 2004).* Vienna (Austria) : Österreichische Gesellschaft für Artificial Intelligence (OEGAI), pp. 613–618.

Bobicev, V., Maxim, V., Prodan, T., Burciu, N., Angheluș, V. (2010): Emotions in Words: Developing a Multilingual WordNet-Affect. In: Gelbukh, A. (ed.) *Computational Linguistics and Intelligent Text Processing (Proceeding of the 11th International Conference CICLing 2010, Iași, Romania).* Heidelberg : Springer, pp. 375-384. (ISBN 978-3-642-12115-9.)

Bond, F., Hitoshi, I., Sanae, F., Kiyotaka, U., Takayuki, K., Kyoko, K. (2009): Enhancing the Japanese WordNet. In: Riza, H., Sornlertlamvanich, V. (eds.) *Proceedings of the 7th Workshop on Asian Language Resources.* Singapore : Association for Computational Linguistics, pp. 1–8.

Das, D., Bandyopadhyay, S. (2009): Word to Sentence Level Emotion Tagging for Bengali Blogs. In: *Proceedings of Short Papers of ACL-IJCNLP.* Singapore : Suntec, pp. 149–152.

Das, D., Bandyopadhyay, S. (2010): Developing Bengali WordNet Affect for Analyzing Emotion. In: *23rd International Conference on the Computer Processing of Oriental Languages.* California (USA) : [s. n.], pp. 35-40.

Ekman, P. (1992): An argument for basic emotions. *Cognition and Emotion*, Vol. 6, No. 3–4, pp. 169–200.

Esuli, A., Sebastiani, F. (2006): SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. In: *Proceedings of the Language Resources and Evaluation Campaign.* Genoa (Italy) : European Language Resources Association (ELRA), pp. 417–422.

Hatzivassiloglou, V., McKeown K. R. (1993): Predicting the semantic orientation of adjectives. In: *Proceedings of the 35th Annual Meeting of the ACL and the 8th Conference of the European Chapter of the ACL.* Stroudsburg (Philadelphia, USA) : Association for Computational Linguistics, pp. 174–181.

Lin K. H.-Y., Yang C., Chen H.-H. (2007): What emotions news articles trigger in their readers? In: Kraaij, W., de Vries, A. P., Clarke, Ch. L. A., Fuhr, N., Kando, N. (eds.) *SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information* Amsterdam : ACM, pp. 733-734. (ISBN 978-1-59593-597-7.)

Miller, A. G. (1995): WordNet: A lexical database for English. *Communications of the ACM*, Vol. 38, No. 11, November, pp. 39-41.

Mohammad, S., Dorr, B. J., Hirst, G. (2008): Computing Word-Pair Antonymy. In: *Proceedings of Empirical Methods in Natural Language Processing and Computational Natural Language Learning Hawaii.* Stroudsburg, PA (USA) : Association for Computational Linguistics (ACL), pp. 982-991.

Mohammad, S., Turney, P. D. (2010): Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In: *CAAGET '10 : Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text.* Los Angeles (CA) : Association for Computational Linguistics, pp. 26–34.

Neviarouskaya, A., Prendinger, H., Ishizuka, M. (2009): SentiFul: Generating a Reliable Lexicon for Sentiment Analysis. In: *Proceedings : 3rd International Conference on Affective Computing and Intelligent Interaction (ACII 2009).* Amsterdam : IEEE, pp. 363-368. (ISBN 978-1-4244-4800-5.)

Pang, B., Lee, L. (2008): Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, Vol. 2, No. 1–2, pp. 1–135. (ISSN 1554-0669)

Quan, C., Ren, F. (2009): Construction of a Blog Emotion Corpus for Chinese Emotional Expression Analysis. In: *Proceedings of the Empirical Method in Natural Language Processing and Association for Computational Linguistics.* Singapore : Association for Computational Linguistics, pp. 1446–1454.

Salovey, P., Mayer, J. (1990): Emotional Intelligence. *Imagination, Cognition and Personality*, Vol. 9, No. 3, pp. 185–211.

Sood, S., Vasserman, L. (2009): ESSE: Exploring Mood on the Web. In: *Proceedings of the 3rd International AAAI Conference on Weblogs and Social Media (ICWSM) Data Challenge Workshop*. 8 pp.

Strapparava, C., Mihalcea, R. (2007): SemEval-2007 Task 14: Affective Text. In: *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007).* Prague: Association for Computational Linguistics, pp. 70–74.

Strapparava, C., Valitutti, A. (2004): Wordnet-affect: an affective extension of wordnet. In: *Proceedings of the 4th International Conference on Language Resources and Evaluation.* Paris: ELRA – European Language Resources Association, pp. 1083–1086. (ISBN 2-9517408-1-6.)

Takamura, H., Inui, T., Okumura, M. (2005): Extracting Semantic Orientations of Words using Spin Model. In: *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics.* Stroudsburg (Philadelphia, USA) : Association for Computational Linguistics, pp. 133–140.

Turney, P. D. (2002): Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL).* Stroudsburg (Philadelphia, USA) : Association for Computational Linguistics, pp. 417–424.

Voll, K. D., Taboada, M. (2007): Not All Words are Created Equal: Extracting Semantic Orientation as a Function of Adjective Relevance. In: Orgun, M. A., Thornton, J. (eds.) *Australian Conference on Artificial Intelligence.* Heidelberg: Springer, pp. 337–346. (ISBN 978-3-540-76926-2.)

Wiebe, J., Riloff, E. (2006): Creating Subjective and Objective Sentence Classifiers from Unannotated Texts. In: *International Conference on Intelligent Text Processing and Computational Linguistics*.

Wiebe, J., Wilson, T., Cardie, C. (2005): Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, Vol. 39, No. 2–3, pp. 165–210.

The work reported in this paper is supported by a grant from the India-Japan Cooperative Programme (DST-JST) 2009 Research project entitled "Sentiment Analysis where AI meets Psychology" funded by Department of Science and Technology (DST), Government of India.

Language-specific Features in Multilingual Sentiment Analysis

Taras Zagibalov
Brandwatch, United Kingdom
e-mail: `taras8055@gmail.com`

Katerina Belyatskaya
Siberian Federal University, Russia
e-mail: `e.o.belyatskaya@gmail.com`

John Carroll
University of Sussex, United Kingdom
e-mail: `J.A.Carroll@sussex.ac.uk`

Abstract

We present newly-produced comparable corpora of book reviews in English and Russian. The corpora are comparable in terms of domain, style and size. We are using them for cross-lingual experiments in document-level sentiment classification. Quantitative analyses of the corpora and the language differences they exhibit highlight a number of issues that must be considered when developing systems for automatic sentiment classification. We describe experiments applying a supervised sentiment classification technique to the corpora. The results of the experiments suggest that differences in the basic characteristics of the two languages and the ways in which sentiment is expressed in the languages lead to significant differences in sentiment classification accuracy.

Key words Sentiment analysis; comparable sentiment corpora

Introduction

We investigate the effect of language-specific features on automated analysis of sentiment in English and Russian. Sentiment analysis is concerned not with the topic or factual content in a document, but rather with the opinion expressed in it. Sentiment analysis has often been broken down into a set of subtasks, including subjectivity classification, opinion classification (or sentiment classification), opinion holder and opinion target extraction, and feature-based opinion mining.

Sentiment classification is usually framed as a two-way classification into positive and negative sentiment, and has been applied at various levels: phrases, sentences, documents and collections of documents. An opinion may have a holder (a person or a group that expresses an opinion) and a target (an object which is being discussed or evaluated). Feature-based opinion mining tries to find opinions about particular features of a product or service (as opposed to an overall opinion about something). Automatic classification of document sentiment (and more generally extraction of opinion from text) has recently attracted much interest. One of the main reasons for this is the importance of such information to companies, other organisations and individuals.

Applications include computer-based tools that help a company see market or media reaction towards their brands, products or services. Another type of application is a search engine that helps potential purchasers make an informed choice of a product they want to buy. Such search engines may include a sentiment classification subsystem that not only presents to a customer the overall sentiment about a product, but may also select positive or negative reviews to illustrate the perceived strengths and weaknesses of a product.

Automated sentiment analysis provides a range of possibilities for researchers in humanities whose studies involve analysis of large amounts of human-generated data. For example, in media studies one might be interested to see whether sentiment regarding the same events is shared in the mainstream media and in social media. Analysis of user-generated content may be very helpful in political studies. For example, the monitoring of political debates in social media may help to estimate the prospects of political candidates in elections or evaluate the effectiveness of political campaigns. The study of 'the language of hatred' contributes to efforts against political and religious extremism and intolerance. Many aspects of social studies may benefit from automatic analysis of sentiments expressed by people in ever-growing social networks. This approach offers unobtrusive and fast access to large amounts of information.

A major current challenge is to be able to automatically extract sentiment information from a variety of documents in different languages. In a recent white paper addressing the role of sentiment analysis in organisations, Grimes (2010) noted 'one axiom of full-circle sentiment analysis is ability to use all relevant sentiment sources'. This obviously includes sources representing different genres and styles, and written in different languages.

The most widely used approach to opinion and subjectivity classification is based on supervised machine learning, in which a system learns from human-annotated training data how to classify documents (e.g. Pang et al., 2002). However, a major obstacle for automatic classification of sentiment and subjectivity is the cost of collecting annotated training data. The rapid growth in the number of languages represented on the Internet and the emergence of new forms of social media makes it increasingly difficult to create and maintain suitable annotated corpora. Rule-based or dictionary-based approaches to sentiment analysis have similar limitations since they rely on large sets of manually created resources that need to match the language data being processed.

There are a number of publicly available sentiment-annotated corpora, such as MPQA (Wiebe et al., 2005), and the Pang and Lee (2004) Movie Review corpus. However, most of these corpora consist only of English text. There are some corpora designed for cross-lingual evaluations, but these seem not to be publicly available, for example the NTCIR MOAT corpora of English, Japanese and Chinese (Seki et al., 2008).

There has been little previous work on applying sentiment analysis to languages with scarce relevant language resources. A notable exception is the work towards producing cross-lingual subjectivity analysis resources from English data by Mihalcea et al. (2007). They use a parallel corpus to adjust a subjectivity lexicon

Table1: Case forms of the Russian adjective *хороший* (good).

Cases	Masculine singular	Feminine singular	Neuter singular	Plural
Nominative	*хороший*	*хорошая*	*хорошее*	*хорошие*
Genitive	*хорошего*	*хорошую*	*хорошего*	*хороших*
Dative	*хорошему*	*хорошей*	*хорошему*	*хорошим*
Accusative	*хорошего / хороший*	*хорошую*	*хорошее*	*хороших / хорошие*
Ablative	*хорошим*	*хорошей*	*хорошим*	*хорошими*
Prepositional	*хорошем*	*хорошей*	*хорошем*	*хороших*

translated from English to Romanian. Other multilingual opinion mining work (in English, Japanese and Chinese) was carried out by Zagibalov and Carroll (2008, 2009), using techniques requiring limited manual input to classify newswire documents with respect to subjectivity and to extract opinion holders and targets.

A related issue that has also received little attention to date is the impact on the design and performance of sentiment analysis systems across languages, stemming from differences in the characteristics of the languages and the means commonly used to express sentiment in them. To address this issue, we have designed and built comparable corpora of book reviews in English and Russian. The corpora are comparable in terms of domain, style and size. The Russian corpus is probably the first sentiment-annotated resource in that language. In the following sections we outline characteristics of the two languages, describe the corpora and quantify their various relevant aspects, and analyse some important language-specific issues that would be likely to impact on automatic sentiment processing. We go on to apply supervised and unsupervised sentiment classification techniques to the corpora to quantify the impact of these language-specific issues on classification accuracy.

Language Characteristics

In this study we focus on English and Russian.

Russian has a relatively complex morphology that comprises gender, case and number forms of adjectives and nouns as well as inclination and tenses, and aspect forms of verbs. For example, the adjective *хороший* (*good*) has the following forms:

- *хороший* – masculine, singular
- *хорошая* – feminine, singular
- *хорошее* – neuter, singular
- *хорошие* – plural (the same for all genders)

Each of these forms may be used with different cases, many of which have different endings (see Table 1).

There are also comparative and superlative forms of the adjective: *лучше* and *наилучший / самый лучший* (the latter is an analytical superlative form). The word can also be used in a short form: *хорош*. The number of forms (16 distinct forms) suggests the need for language-specific lexical processing (for example with a morphological analysis tool) before any application-level processing could take place.

English uses morphological means to express grammatical tense and aspect for verbs, and singular and plural for nouns. Arguably the most important part of speech for sentiment analysis – adjectives – also have comparative and superlative forms which sometimes are formed irregularly (e.g., *good – better – best* and *bad – worse – worst*). Nevertheless, the variation of grammatical forms in English is not as complex as in Russian.

In a language-based application, such as sentiment analysis, without lexical processing (such as morphological analysis, stemming or lemmatisation) one may have the problem of data sparseness since numerous word forms would 'hide' a single word, even if a large amount of corpus data was available. However, lexical processing of this type is necessarily language-dependent, making it expensive to use this type of approach in a system that covers multiple languages.

The Corpora

Corpora Content

Our English and Russian book review corpora consist of reader reviews of science fiction and fantasy books by popular authors. The reviews were written in 2007, so the language used is current.

The Russian corpus consists of reviews of Russian translations of books by popular science fiction and fantasy authors, such as S. King, S. Lem, J. K. Rowling, T. Pratchett, R. Salvatore, J. R. R. Tolkien as well as by Russian authors of the genre such as S. Lukyanenko, M. Semenova and others. The reviews were published on the website `www.fenzin.org`.

The English corpus comprises reviews of books by the same authors, if available. If some of the authors were not reviewed on the site or did not have enough reviews, they were substituted with other writers of the same genre. As a result, the English corpus contains reviews of books such as: S. Erickson (*Guardians of the Moon, Memories of Ice*), S. King (*Christine, Duma Key, Gerald's Game, Different Season* and others), S. Lem (*Solaris, Star Diaries of Iyon Tichy, The Cybriad*), A. Rise (*Interview with the Vampire, The Tale of the Body Thief* and others), J.K. Rowling (*Harry Potter*), J. R. R. Tolkien (*The Hobbit, The Lord of the Rings, The Silmarillion*), S. Lukyanenko (*The Night Watch, The Day Watch, The Twilight Watch, The Last Watch*), and a few others. The reviews were published on the website `www.amazon.co.uk`.

Although both of the sites from which the reviews were collected feature review-ranking systems (e.g., one to ten stars), many reviewers did not use the system or did not use it properly. For this reason all of the reviews were read through and hand-annotated. There were a lot of reoccurring short reviews such as: Хорошо (*Good*); Интересная книга (*Interesting book*); Супер! (*Superb!*); Нудятина!! (*Boring!!*); Ниже среднего (*Below average*); *Awesome!*; *Amazing!*; *The best book I've ever read!*; *Boring*, and so on. These reviews were added to the corpus only once. Also both sites had a number of documents which did not have any direct relation to book reviewing, such as advertisements, announcements and off-topic postings. Such texts were excluded as irrelevant. The documents that were included in the corpora were not edited or altered in any other way.

We annotated each review as 'POS' if positive sentiment prevails or 'NEG' if the review is mostly negative based on the tags assigned by reviewers, but moderated where the tag was obviously incorrect. Each corpus consists of 1,500 reviews,

Table2: Overall quantitative measures of the English and Russian corpora.

	Mean tokens POS	Mean tokens NEG	Total types POS	Total types NEG
English	58	58	7349	8014
Russian	30	38	9290	12309

half of which are positive and half negative. The annotation is simple and encodes only the overall sentiment of a review, for example:

> [TEXT = POS]
> Hope you love this book as much as I did. I thought it was wonderful!
> [/TEXT]

The English reviews contain a mean of 58 words (the mean length for positive and negative reviews being almost the same). Positive Russian reviews have a mean length of only 30 words; negative reviews are slightly longer, at 38 words (see Table 2). It is not possible to compare these figures directly between the languages as they have different grammar structures which makes English more 'wordy', as it has function words (articles, auxiliary verbs) which are almost completely absent in Russian.

As noted above, Russian, being a synthetic language, has many forms of the same lemma. This results in a large number of distinct word forms: the corpus contains a total of 13 472 word forms, with 6 589 (42%) in positive reviews and 8 993 (58%) in negative. The total number of words in the corpus is 50 745, which means that every word form was used a little more than three times on average. The English corpus has only 7 489 distinct word forms in the whole corpus, 4 561 (47%) in positive reviews, and 5 098 (53%) in negative. These figures also suggest that Russian reviewers used a richer vocabulary for expressing *negative* opinions (compared to the number of unique words used in Russian positive reviews) than English reviewers.

Further evidence of the different ways in which people distinguish sentiment polarity in Russian compared with English is the distribution of the lengths of positive and negative reviews. The Russian corpus has a large number of short reviews (less than 50 words) with a median of 15 words for positive reviews and 10 words for negative reviews. Apart from the language-specific differences mentioned above that partly account for the smaller number of words in Russian documents, there is a clear difference from English reviews in terms of length. The English reviews feature a more or less equal number of documents of different lengths (mostly in the range 15 to 75 words). The prevalence of short reviews in the Russian corpus, together with the rich morphological variation, may lead to data sparseness which would be a problem for current sentiment classification techniques.

Ways of Expressing Sentiment

Sentiment can be expressed at different levels in a language: from lexical and phonetic levels up to discourse level. This range is reflected in the corpora (see Tables 3

Figure1: Distribution of documents by number of words

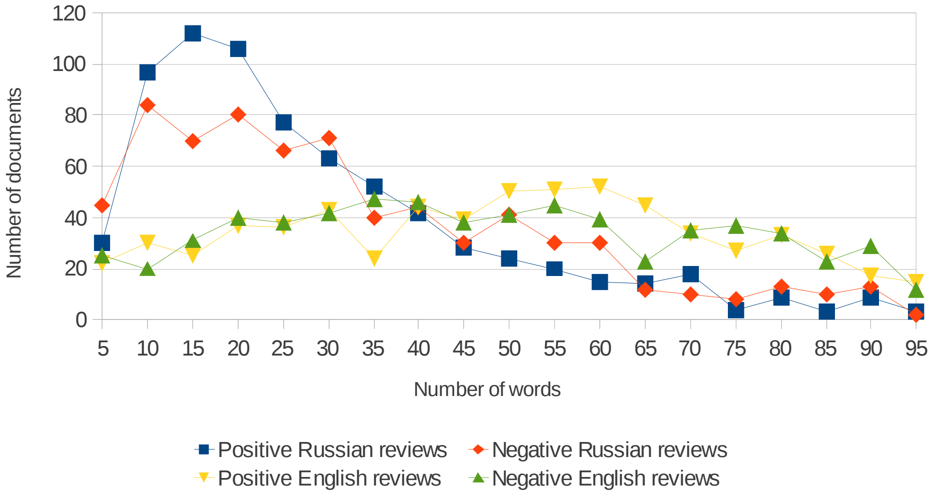


Table3: Ways of expressing sentiment in the English Book Review Corpus (numbers of documents)

	Syntactic	Lexical				Phonetic
		Verb	Adjective	Noun	Other	
Positive	432	312	708	225	325	12
Negative	367	389	652	238	407	16
Total	799	701	1360	463	732	28

and 4)[1]. As the Tables show, the authors of reviews in the two languages express sentiment in slightly different ways. In English they make heavy use of adjectives to express sentiment (this class of words is used to express sentiment in a third of all documents). In contrast, in Russian they use verbs as often as adjectives to express sentiment (both of these classes are used in about a quarter of all reviews) and make more use of nouns (expressing sentiment in 15% of all documents compared to 11% in English). The Russian corpus also demonstrates a tendency to combine different ways of expressing sentiments in a document: the total number of uses of different ways in the English corpus is 4,083 compared to 4,716 in Russian, which means that given an equal number of reviews for each language, Russian reviews tend to have more ways of expressing sentiment per document.

Lexical Level

Adjectives are the most frequent way of expressing opinions in both corpora, closely followed by verbs in the Russian corpus. 1,215 Russian reviews use adjectives to ex-

[1] All the numerical data presented below comes from manual counting and is not represented in the corpus annotation.

Table4: Ways of expressing sentiment in the Russian Book Review Corpus (numbers of documents).

	Syntactic	Lexical				Phonetic
		Verb	Adjective	Noun	Other	
Positive	417	492	648	374	367	27
Negative	475	578	567	334	394	43
Total	892	1070	1215	708	761	70

press sentiment and 1,070 reviews use verbs. In the English corpus there are 1,360 reviews that use adjectives, but only 701 use verbs to express opinion.

Apart from adjectives, which are recognised as the main means of expressing evaluation, other parts of speech are also often used in this function, most notably verbs and nouns. The English reviews also feature adverbials, and both languages also use interjections.

Akimova and Maslennikova (1987) observe that opinions delivered by means of verbs are more expressive compared to opinions expressed in other ways. This is explained by the fact that a verb's denotation is a situation and the semantic structure of the verb reflects linguistically relevant elements of the situation described by the verb. Verbs of appraisal not only name an action, but also express a subject's attitude to an event or fact.

Consider the following examples:

- *I truly loved this book, and I KNOW you will, too!*
- *понравилось, научная фантастика в хорошем исполнении*

  I liked it, it's science fiction in a very good implementation

The English verbs *loved* and *liked* describe an entire situation which is completed by the time of reporting it. This means that a subsequent shift in sentiment polarity is all but impossible:

- **I truly loved this book, but it turned out to be boring.*

However, adjectives usually describe only attributes of certain members of a situation leaving a significant amount of context aside:

- *The story is pretty good but it stretches on and on.*

In the example above a positive sentiment towards the story is shifted to negative. A verb is less usual in such a context:

- (?) *I liked the story but it stretches on and on.*

Nouns can both identify an object and provide some evaluation of it. But nouns are less frequently used for expressing opinion compared to verbs. Nonetheless in the Russian corpus, nouns were used more than in the English corpus. There are 708 Russian reviews that have opinions expressed by nouns, however only 463 English reviews made use of a noun to describe opinion. The most frequent such nouns used in Russian reviews are чудо (*miracle*), классика (*classics*), шедевр (*masterpiece*), гений (*genius*), прелесть (*delight*), бред (*nonsense*), мура (*raspberry*), жвачка (*mind-numbing stuff*), ерунда (*bugger*).

Phonetic Level

Although the corpora consist of written text and do not have any speech-related mark-up, some of the review authors used speech-related methods to express sentiment, for example:

- *This was a sloooow, frail story*
- *A BIG FAT ZEEROOOOOOOOOOOOO for M.A*
- *i have to say is a good boooooooooooooooooooook!*
- ***Ну что сказать...чепуха...ЧЕ-ПУ-ХА.***
  What should I say... boloney... BO-LO-NEY
- ***Ндааааа.............такую муть давно не видел***
  Weeeeelll........ I haven't seen such a stinkaroo for long
- ***абалденная книшкаааааа!!!!!!!!!!!!!!!))) оч давно её люблю))***
  jaw-droppin' boooooook!!!!!!!!!!!!!!!))) been lovin' it for long
- ***Мозг ломиться от этого несоответствия... и получает ооочень большой кайф!!!***
  My brain is bursting because of this inconstancy... and it enjoys it veeery much!!!
- ***Читать ВСЕЕЕЕЕЕЕЕЕЕМ***
  Read, EVERYBOOOOODY

Another way to express opinion in Russian is based on the use of a sub-culture language, Padonky. This sociolect has distinctive phonetic and lexical features that are distant from 'standard' Russian (both official and colloquial). For example, a phrase usually used to express a negative attitude to an author about his book:

- ***Аффтор, выпей ЙАДУ***
  (lit) Autor, drink some POIZON

Padonky is close to some variants of slang (corresponding in English to expressions such as *u woz*, *c u soon* etc.), however it is more consistent and is used quite often on the Web.

Sentence Level

Sentence-level means of expressing sentiment (mostly exclamatory clauses, imperatives or rhetorical questions) is slightly more frequent in the Russian corpus than in the English: 892 and 799 respectively. The distribution of positive and negative sentiments realised at the sentence level is opposite in the two corpora: syntactic means are used more frequently in negative reviews in Russian but they are more frequent in positive reviews in English.

One particularly common sentiment-relevant sentence-level phenomenon is the rhetorical question. This is a question only in form, since it usually expresses a statement. For example:

- ***Иоткуда столько восторженных отзывов? Коробит от крутости главных героев***
  Why are there so many appreciative reviews? The 'coolness' of the main characters makes me sick

- *Что же такого пил/принимал/нюхал автор, чтобы написать такое?*
  What did the author drink / eat / sniff to write stuff like that?
- *Интересно, кто-нибудь дотянул хотя бы до середины? Лично я - нет.*
  I wonder if anyone managed to get to the middle? I failed.

Considering imperatives, the review author is telling their audience 'what to do', which is often to read a book or to avoid doing so.

- *Run away! Run away!*
- *Pick up any Pratchett novel with Rincewind and re-read it rather than buying this one*
- *Читать однозначно.*
  Definitely should read.
- *Читать !!!!!!!!!! ВСЕМ*
  Read!!!!!!!! EVERYONE

Another way of expressing sentiment through syntactic structure is by means of exclamatory clauses, which are, by their very nature, affective. This type of sentence is widely represented in both corpora.

- *It certainly leaves you hungering for more!*
- *Buy at your peril. Mine's in the bin!*

Discourse Level

Some means of sentiment expression are quite complex and difficult to analyse automatically:

- *Иэто автор вычислителя и леммингов? ... НЕ ВЕРЮ! Садись, Громов, два.*
  (lit) So this author calculator and lemmings? ... (DO)NOT BELIEVE! sit, gromov, two.
  So is this the author of The Calculator and of The Lemmings? ... Can't believe it! Sit down, Gromov, mark 'D'!

This short review of a new book by Gromov, the author of the popular novels *The Calculator* and *The Lemmings*, consists of a rhetorical question, an exclamatory phrase and an imperative. All of these means of expression are difficult to process. Even the explicit appraisal expressed by utilising a secondary school grade system is problematic as it requires specialised real-word knowledge about the meaning of the numeral 'two'[2] in this context.

The example below also features an imperative sentence that is used to express negative sentiment. This review also lacks any explicit sentiment markers. The negative appraisal is expressed by the verbs *stab* and *burn* which only in this context show a negative attitude.

- *Stab the book and burn it!*

[2] Russian schools use a 5-point marking system, with 5 as the highest mark. Thus a '2' can be considered as equivalent to a 'D'.

Discussion

The reviews in English and in Russian often use different means of expressing sentiment, many of which are difficult (if at all possible) to process automatically. Often opinions are described through adjectives (86% of reviews contain adjectives). The second most frequent way of expressing sentiment is through verbs (59% of reviews have sentiment-bearing verbs). Less frequent is the noun, in 39% of reviews. Sentence-level and discourse-level sentiment phenomena are found in 56% of reviews. 3% of reviews contain phonetic sentiment phenomena.

Issues that may Affect Sentiment Analysis

One of the features of web content not mentioned above is a high level of mistakes and typos. Sometimes authors do not observe the standard rules on purpose (for example using sociolects, as outlined above). For example, in the corpora 52% of all documents contain spelling mistakes in words that have sentiment-related meaning. The English corpus is less affected as authors do not often change spelling on purpose and use contractions that have already become conventional (e.g., *wanna, gonna* and *u*). However, the number of spelling mistakes is still high: 48% of reviews contain mistakes in sentiment-bearing words. The proportion of misspelled words in the Russian corpus is higher, at 58%.

Of course, a spelling error is not always fatal for automatic sentiment classification of a document, since reviews usually have more sentiment indicators than just one word. However, as many as 8% of the reviews in both corpora have all of their sentiment-bearing words misspelled. This would pose severe difficulties for automatic sentiment classification.

Another obstacle that makes sentiment analysis difficult is topic shift, in which the majority of a review describes a different object and compares it to the item under review. The negative review below is an example of this:

- ***Дочитала с трудом. Ничего интересного с точки зрения информации. Образец интеллектуального детектива – романы У.Эко. И читать приятно, и глубина философии, и в историческом плане познавательно. А в эстетическом отношении вообще выше всяких похвал.***

  Hardly managed to read to the end. Nothing interesting from the point of view of information. An example of intellectual detective stories are novels by U.Eko. It's a pleasure to read them, and (they have) deep philosophy, and are quite informative from the point of view of history. And as for aesthetics it's just beyond praise.

The novel being reviewed is not the one being described, and all the praise goes to novels by another author. None of the positive vocabulary has anything to do with the overall sentiment of the review's author towards the book under review.

Other reviews that are difficult to classify are those that describe some positive or negative aspects of a reviewed item, but in the end give an overall sentiment of the opposite direction. Consider the following positive review:

- ***Сюжет довольно обычен, язык изложения прост до безобразия. Много грязи, много крови и смерти. Слишком реально для сказки коей является фэнтези. Но иногда такие книги читать полезно, ибо они описывают неприглядную реальность.***

> *The plot is quite usual, the language is wickedly simple. A lot of filth, a lot of blood and death. Too true-to-life for a fairy-tale, which a fantasy genre actually is. But it is useful to read such books from time to time, as they depict ugly reality.*

The large number of negative lexical units may mislead an automatic classifier to a conclusion that the review is negative.

The three issues described above are present in approximately one-third of all reviews in the corpora. This suggests that a sentiment classifier using words as features could only correctly classify around 55–60% of all reviews.

This performance may be even worse for the Russian corpus since many of its reviews feature very unexpected ways of expressing opinion. Unlike most of the English reviews, in which a reviewer simply gives a positive or negative appraisal of a book, backing it with some reasoning and probably providing some description and analysis of the plot, Russian reviews often contain irony, jokes, and use non-standard words and phrases, making use of a variety of language tools, as illustrated in the following examples:

- *Скушнаа. дошёл до бегства ГГ в мир Януса, и внезапно понял (я), что гори он (ГГ) хоть синим пламенем*
  Booorin'. got to the (episode of) GG fleeing to the world of Janus, and suddenly (I) realised that let it (GG) burn with blue flames (I do not at all care about GG)

- *Я эту муть не покупал. Shift+del.*
  I didn't buy this garbage. Shift+del.

Since there are more reviews of this kind in the Russian corpus than in the English, it is very likely that a Russian sentiment classifier would have lower accuracy.

Sentiment Classification Experiments

In this section we apply supervised sentiment classification techniques to the English and Russian corpora to quantify the impact on accuracy of the language-specific issues discussed above.

Feature Extraction

Approaches to sentiment classification of documents using machine learning require a set of features to be extracted from each document. Most work on English uses word forms as the features, tokenised by splitting the character stream at whitespace and punctuation characters (e.g., Pang et al., 2002).

An alternative approach is to use 'lexical units' as features, where a lexical unit is any commonly-occurring sequence of characters, which may constitute a part of a word, a complete word or even a phrase. This approach avoids the need for word segmentation, and can also capture some grammatical and syntactic information, because lexical units can incorporate function words and parts of grammatical constructions. We extracted lexical units in a pre-processing step by finding the longest strings occurring at least twice in the corpus.

The English book review corpus produced 7,913 such lexical units. Some of these are word sequences expressing features that are often discussed by reviewers,

Table5: Classification results (10-fold cross-validation, words)

	NBm			SVM		
	Precision	Recall	F-measure	Precision	Recall	F-measure
English	0.85	0.85	0.85	0.83	0.83	0.83
Russian	0.78	0.78	0.78	0.73	0.73	0.73

such as *the plot* or *the characters*, as well as phrases that are used for appraisal such as *good performance* and *best performance*.

The same approach was applied to the Russian corpus; despite the language's complex morphology one might expect the technique to be able to capture more unchangeable (stable) units as well as frequent word forms. This indeed turns out to be the case, since the approach extracts some 'semi-stemmed' forms that comprise the most important part of the word, leaving out affixes denoting minor grammatical features, for example, the lexical unit **бессмыленн** which is a common part of the word forms **бессмыленный, бессмыленная, бессмыленных, бессмыленного** and many others meaning *senseless*. The Russian corpus produced 8,372 lexical units.

Results

We used two machine learning algorithms, Naïve Bayes multinomial (NBm) and Support Vector Machines (SVM)[3], trained and evaluated on the corpora of English and Russian, and the two techniques for feature extraction (word forms and lexical units). The evaluations used 10-fold cross-validation.

With word forms as features, in order to make the resulting lexicons comparable in terms of their elements' frequencies we filtered out all words that occurred less than 10 times. We extracted all words from the corpora but did not process them in any way. 1,075 words were extracted from the Russian corpus and 1,247 words from the English book reviews. The classification results are shown in Table 5. The results for Russian are much worse than for English, which might be expected since the abundance of word forms in Russian makes the data sparse.

We also ran the same machine learning algorithms with lexical units extracted from the two corpora as features. The results are shown in Table 6. It could be expected that the 'semi-stemming' property of lexical units would even out differences in accuracy due to different levels of morphological productivity in the two languages. Indeed, the accuracies for Russian are much improved over using word forms as features. Nevertheless, the accuracies for Russian are still lower than for English; this might be explained by the apparently more diverse means of expressing opinion in the Russian corpus than the English one, as discussed above.

Conclusions

In this paper we presented comparable corpora of English and Russian book reviews, examined language-specific features of the reviews that are relevant to senti-

[3] We used WEKA 3.4.11 (`http://www.cs.waikato.ac.nz/~ml/weka`)

Table6: Classification results (10-fold cross-validation, lexical units).

	NBm			SVM		
	Precision	Recall	F-measure	Precision	Recall	F-measure
English	0.88	0.88	0.88	0.84	0.84	0.84
Russian	0.81	0.81	0.81	0.78	0.78	0.78

ment classification, and showed that sentiment in different languages is expressed in slightly different ways, covering all levels of the language: from phonetic to discourse.

We also considered features of the languages themselves; in particular, the complex morphology of Russian may affect the performance of a supervised classifier that does not use any pre-processing techniques, such as stemming or lemmatisation. However, an approach based on identifying common 'lexical units' in a pre-processing step performed much better on the Russian corpus compared to using words as features.

We also found significant differences in sentiment classification accuracy between English and Russian, despite using comparable corpora for training and testing. We conclude that more work is needed to determine the best approach to sentiment analysis for different languages.

References

Akimova, T., Maslennikova, A. (1987): *Lingvisticheskie issledovanija.* Moscow : [s. n.], pp. 3–33. Chapter: Imperativa i Ocenka (Semantics of Imperatives and Appraisal).

Grimes, S. (2010): The Three Secrets to Successful Sentiment Analysis [on-line]. [cit. 2011-09-13]. Retrieved February 16, 2010. Available at: `http://www.mycustomer.com/topic/customer-intelligence/seth-grimes-how-get-sentiment-analysis-right/103102`.

Mihalcea, R., Banea, C., Wiebe, J. (2007): Learning multilingual subjective language via cross-lingual projections. In: *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics.* Prague, Czech Republic : Association of Computational Linguistics, pp. 976–983.

Pang, B., Lee, L., Vaithyanathan, S. (2002): Thumbs up?: sentiment classification using machine learning techniques. In: *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing (Vol. 10).* Stroudsburg (Philadelphia, USA) : Association for Computational Linguistics, pp. 79–86.

Pang, B., Lee, L. (2004): A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In: *Proceedings of the 42nd Annual Meeting of the Association of Computational Linguistics.* Stroudsburg (Philadelphia, USA) : Association for Computational Linguistics, pp. 271–278.

Seki, Y. (2008): A multilingual polarity classification method using multi-label classification technique based on corpus analysis. In: *Proceedings of the NTCIR-7 MOAT Workshop Meeting.* Tokyo (Japan) : National Institute of Informatics, pp. 284–291.

Wiebe, J., Wilson, T. A., Cardie, C. (2005): Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2), pp. 165–210.

Zagibalov, T., Carroll, J. (2008): Almost unsupervised cross language opinion analysis at NTCIR-7. In: *Proceedings of the NTCIR-7 MOAT Workshop Meeting.* Tokyo (Japan) : National Institute of Informatics, pp. 204–209.

Zagibalov, T., Carroll, J. (2009): Multilingual opinion holder and target extraction using knowledge-poor techniques [on-line]. [cit. 2011-09-13]. Available at: `http://www.informatics.sussex.ac.uk/research/groups/nlp/carroll/papers/ltc09.pdf`.

Acknowledgement

The first author was supported by the Ford Foundation International Fellowships Program.