A Study on Life Aspect Inference based on Association with Latent Topics

Shuhei Yamamoto

Graduate School of Library, Information and Media Studies

University of Tsukuba

March 2016

A Study on Life Aspect Inference based on Association with Latent Topics

Shuhei Yamamoto

Many information sharing services currently exist, such as community knowledge sharing sites, blogs, and micro-blogs. Twitter, which is one of the most popular social media services, had 280 million monthly active users at the end of the September 2013. Since it only permits users to post short passages up to 140 characters, users can easily share their experiences and opinions about daily events. Twitter posts are often both useful and timely because they typically comment on current events. For example, tweets about traffic jams or accidents are quite valuable for users who will travel past those places. Supermarket sales and bargain information are also helpful for neighborhood consumers. Such tweets, which are highly regional, up-to-date, and beneficial to others. We call such tweets real life tweets.

Information is used in various aspects of life. Real life tweets can accommodate such aspects. For example, such tweets as "My train is late!" are categorized as a "Traffic" aspect and support train commuters. Such posts as "Today, bargain sale items are 50% off!" are categorized as an "Expense" aspect and support shoppers. For presenting real life tweets based on user contexts, we classify them into 14 aspects, which are assumed to be the life aspects of users, based on the Yahoo directory, "local community", and "life" in the Japanese version of Wikipedia.

Users post various types of tweets. "Nods" and sympathetic phrases frequently appear on Twitter. For example, "Thank you" and "I see" often appear in posts. These posts do not directly support the real life situations of other users. We believe that users want a method of locating beneficial tweets on Twitter. Such nods and sympathies simply impede the discovery of substantive tweets.

Users generally expect contents that reflect their particular interests. However, the above problems occurr in many applications and social media. For example, many

spam e-mails are delivered daily. Some blogs offer no interesting content, and news articles mention various categories. These hinder not only the access to information by users but also bury interesting contents under a flood of excessive information. Such problems can be solved by estimating the labels of each bit of content so that users can rapidly access contents based on labels of interests.

In this study, we estimate the aspects of unknown tweets by addressing a supervised machine learning approach that trains a model by labeled data to estimate unknown data using it. To estimate the aspects of unknown tweets, we faced these four challenges:

1. We must estimate the aspects of unknown tweets with just a few feature terms because 45, which is the average number of characters in tweets, is shorter than general documents.

2. We must achieve the highest accuracy by the latest labeled tweets because the lives of people change quickly and new terms might appear. Annotating the aspects of many tweets is difficult because the aspects are hand-labeled, and so we need to estimate them by a small set of labeled tweets.

3. Depending on the tweets, we have to estimate several aspects of a tweet. For example, the following tweet "A heavy snowstorm caused a traffic accident near the JFK airport," mentions a snowstorm and a traffic accident. Its main topic is the accident, but it also provides weather information. Therefore, we multi-label it as both Traffic and Weather.

4. An approach that estimates several aspects of a tweet can clearly provide real life information for specific users. On the other hand, exhaustive-oriented users might expect broad information that includes the specific aspects. In other words, accuracy-oriented users might desire strictly selected real life information on specific aspects. When we visit sightseeing locations, we want information

about them. The multi-label classification approach fails to achieve such tightly associated aspects with the same weight.

In typical supervised machine learning methods as document classification, naive Bayes classifier and support vector machine (SVM) are widely known. Both methods have been extended to multi-labeling, and labeled-latent Dirichlet allocation (L-LDA) has also been proposed to multi-label documents. These methods address the classification of relatively long documents and achieve high accuracy using sufficient training datasets. However, in the cases of such short documents as tweets and a small set of labeled data, these methods probably cannot achieve adequate performance because tweets have too few feature terms.

In this study, we propose a hierarchical estimation framework (HEF) based on associations between topics and aspects to satisfy the above challenges. Its fundamental idea is composed of both unsupervised and supervised machine learning techniques. In the first phase, it extracts topics from a sea of tweets using latent Dirichlet allocation (LDA). In the second phase, it calculates the relevance among topics and aspects using a small set of labeled tweets to build associations among them. HEF calculates the aspect scores for unknown tweets using the association between topics and aspects based on the terms extracted from the tweets.

Although typical machine learning methods directly calculate the likelihood of terms, HEF calculates the relevance between topics and aspects using a small set of labeled tweets and builds associations based on relevance. One HEF feature expands the terms of the tweet to topics, estimates the contents in the topics, and calculates the aspect scores by associations between topics and aspects. Thus, by two-phase estimation based on topics as the middle-layer, even if a term does not appear as training data, we can calculate the aspect scores using it. In other words, HEF stochastically expands the terms of tweets using topics. Although it inadequately expand terms, estimation performance decrease. Therefore, we preliminarily calculate the occurrence probability of the terms in each topic by LDA using a large amount

of tweets.

In multi-label classification, when the aspect scores exceed a threshold, the aspects are estimated for tweets. HEF introduced entropy feedback mechanisms in the second phase to overcome the problem of competitive associations among aspects. Based on these extensions, the associations between topics and aspects are refined and the estimation precisions are increased. In experimental evaluations using Japanese tweets posted in the Kyoto area, we clarified that HEF appropriately estimates the aspects for unknown tweets. HEF, which introduced entropy feedback, builds refined associations that linked feature topics to each aspect and showed the highest F-measure among typical methods of multi-label classification. With less training data, the precision, recall, and F-measure values of the typical methods rapidly dropped; however, HEF retained its high evaluation values.

The aspect distribution is represented by the probability distribution in each tweet. Accurately inferring the probability distribution of aspects means supporting either the strict or broad associations between tweets and aspects. As an inference approach of probability distribution, we naturally extend HEF by normalizing scores and propose an optimal association building method based on t-test, which is an efficient strategy to manage the relationship between topics and aspects. We assume that the training data are not given as the probability distributions of the aspects based on a training model of a typical classification method. Our challenge in this study is to train from labeled tweets and infer the probability distribution of the aspects of unknown tweets. The experimental evaluations of this study prepared a small set of labeled tweets based on classifications by three examinees and calculated the probability distributions of each tweet from them. In the case of single label training, HEF showed significantly lower JS Divergence and Euclidean Distance values than every baseline method based on sharing topics by several aspects.

From these results, the HEF scheme is an effective life aspect inference method of multi-label classification and probability distribution using a small labeled dataset for

such short sentences as tweets because the associations between topics and aspects appropriately expanded the terms.

潜在トピックとの対応関係に基づく生活の局面推定に関する研究

山本 修平

現在，知識共有コミュニティサイトやブログ，マイクロブログなど，多くの情報共有サービスが存在している．ツイートと呼ばれる短文を投稿する Twitter は，最も広く普及しているマイクロブログの 1 つであり，2013 年 9 月末に 2 億 8000 万を超える月間アクティブユーザ数を記録している．ユーザは自らの経験や意見，また日常生活でのイベントなど，身近な「今」を投稿している．このため，他のユーザにとっても最新かつ有益なツイートが多く，たとえば，電車の遅延情報は交通機関を利用するユーザに役立ち，近所のスーパーマーケットの特売情報は買物に出かけようとしているユーザを支援できる．これらのような地域性が高く新鮮かつ，他のユーザに有益なツイートを本研究では「実生活ツイート」と呼ぶ．

実生活ツイートが実際にユーザを支援した場面として，2011 年 3 月に起きた東日本大震災が挙げられる．地震の直後，震災の被害に遭った地域では断水や食料共有の不足，交通機関の運行停止など，大きな混乱が生じた．その際，給水や食料配布が行われる場所，電車やバスの運行情報について記述された有益なツイートが数多く投稿され，多くの生活者を支援したと報告されている．

実生活ツイートは生活の様々な局面に対応している．たとえば，「電車が来ない」というツイートは生活の中の「交通」の局面に対応し，これから電車に乗ろうとしているユーザを支援できる．「雨が降ってきた」というツイートは「気象」の局面に対応し，これから外出する人や，洗濯しようとする人など，幅広いユーザを支援できる．本研究では，人々の生活を典型的な 14 の局面に整理している． 特定の局面に関するツイートを頻繁に投稿するユーザも多く存在する．たとえば，災害の局面では，Twitter が公式にライフラインに関するツイートを投稿するユーザを地域毎に収集している．東京都交通局は，電車の遅延などの運行情報をつぶさに投稿している．このような公式アカウントをフォローすることにより，ある局面に関して公の情報が得られる．しかし，より局所的な生活情報を素早く取得するためには，日常生活におけるソーシャルセンサとして機能している一般ユーザのツイートも無視できない．

一方で，ユーザは実生活ツイート以外のツイートも数多く投稿している．特に,「ありがとう」や「そうなんだ」といった，誰かの投稿に対する相槌や共感など，ユーザの生活を直接支援しないツイートが多い．このようなツイートの混在は，実生活ツイートの発見を妨げる原因となっている．

一般的に，ユーザは自身にとって興味のある情報の取得を望んでいると考えられる．しかし，これまでに述べた課題は，多くのアプリケーションやソーシャルメディアに起きている．このような課題により，ユーザが必要な情報に即座にアクセスできないだけでなく，必要な情報が埋没してしまう．この解決方法として，各情報に対してユーザに理解できるラベルを推定することが挙げられる．

本研究の目標は，ユーザの所望する特定の局面を提供するために，未知のツイートに局面を推定することである．この目標を達成するための素朴な方法は，人手で各局面に強く関連する単語を列挙し，未知のツイート中に出現する単語と照合することにより，関連度の高い局面を推定することである．このような方法は高い精度を実現できるが，本研究で対象とする実生活ツイートは様々な局面を含んでおり，局面に関連するすべてのキーワードを列挙することは困難である．そこで，本研究では教師あり機械学習に基づくアプローチにより，ツイートに局面を推定することを試みる．ここでの課題は，以下に示す 4 項目である．

1. ツイートは平均 45 文字と短いことから，少ない手がかり語からツイートの言及している局面を推定する必要があること．

2. 人々の生活は時間とともに変化していくことから，なるべく最新に投稿されたツイートを訓練データとすることが望ましく，できる限り少量の訓練データで高い推定精度が得られること．

3. ツイートによっては，複数の局面を推定する必要があること．たとえば,「猛吹雪が原因で，JFK 空港の近くで交通事故が起きました」というツイートは,「猛吹雪」と「交通事故」に言及している．ツイートの主題は「交通事故」であるが，同時に「猛吹雪」という気象情報も提供している．このため，このツイートには「交通」と「気象」の両局面を推定することが相応しい．

4. ツイートに局面のラベルを割り当てることにより，明確な実生活情報をユーザに提供することができる．一方で，ある局面に少しでも関連しているツイートは全て閲覧したい網羅性を重視するユーザや，ある局面に対して正確に言及しているツイートのみ閲覧したい正確性を重視するユーザの存在が考えられる．このような指向を持つユーザに対しては，マルチラベル分類によるアプローチでは対応できない．

従来の教師あり機械学習による手法では，Naive Bayes 分類器や SVM を用いた手法が広く知られている．両手法ともマルチラベリングへ拡張されており，またトピックモデルの 1 つである Labeled LDA もマルチラベリングを目的に提案されている．いずれの手法も十分な訓練データを用いることで，ブログやニュース，Web ページなどの比較的長い文書を分類することを目的とし，高い推定精度を示している．しかし，本論文で課題とする，短文である場合や訓練データが少ない場合には，考慮できる手がかり語が少なくなるため，十分な性能が得られないと考えられる．

本研究では，上記で述べた課題を解決するために，潜在的なトピックと局面の対応関係に基づく階層的推定法を提案する．階層的推定法の基本的なアイデアは，教師なし学習と教師あり学習の両方を組み合わせ，2 段階の学習を行うことにある．第 1 段階では，教師なし学習として知られる潜在的ディリクレ配分法（LDA）を用いて，大量のツイート集合からトピックを抽出する．第 2 段階では，局面ラベルが付与された少量のツイートを用いて，抽出した潜在トピックと局面の関連度を算出し，局面に複数トピックを結びつけた対応関係を構築する．実際に未知のツイートに局面を推定する際は，ツイートに出現する単語から，その単語の出現するトピックの生起確率とそのトピックが対応関係を持つ局面への関連度を用いて，局面毎にスコアを算出する．

従来の教師あり機械学習手法は，訓練データから直接クラスラベルに対する単語の尤度を学習しているが，提案する階層的推定法は．局面とトピックの関連度を算出し，関連度に基づいて対応関係を構築する．提案手法の特徴は，ツイートに出現する単語をトピックに展開し，ツイートが言及している話題をトピックという単位で確率的に拡張した後に，少量の訓練データであらかじめ学習したトピックと局面の関連度からツイートに局面を推定することにある．すなわち，局面を推定しようとするツ

イートに出現する単語を，トピックを使って確率的に拡張していることに特徴がある．しかし，むやみに拡張するとノイズとなる単語によって推定精度が低下することから，LDA によって大量のツイートからトピックに属する単語の出現確率をあらかじめ学習しておく．

マルチラベル分類においては，スコアが閾値を超えた局面をツイートに付与することにより実現する．ここでは，多くの局面が同じトピックに対して結びつく競合問題を解決するために，Entropy Feedback という機構を階層的推定法の第 2 段階に導入する．この拡張に基づき，トピックと局面の対応関係は洗練され，推定性能の向上を狙う．Entropy Feedback は現時点でのトピックと局面の対応関係に対してエントロピーを算出し，その値に基づいてフィードバック係数を求め，再度関連度を算出し直すことによって実現する．京都市内で投稿された日本語ツイートを用いた評価実験の結果，階層的推定法は未知のツイートに適切に局面を付与できることを明らかにした．Entropy Feedback を導入した提案手法は，それぞれの局面に特徴的なトピックが強い関連度で結びついており，実際に対応関係が洗練されたことを確認した．提案手法と従来のマルチラベル分類手法の適合率，再現率，F 値を用いて推定性能を比較した結果，階層的推定法は高い F 値を示した．特に，訓練データの数を減らした場合で，比較手法の推定精度が低下した中で，提案手法はほとんど下降しないという特徴が明らかになった．

ユーザの指向に合わせた実生活ツイートを提供するタスクにおいては，未知のツイートに対して生起する局面の確率分布を推定することで実現する．ここでは，t 検定に基づく最適なトピックと局面の対応関係を構築する手法を提案する．また，本研究ではラベルが付与された訓練データでモデルを学習し，入力された未知のツイートに対しては確率分布を推定する．評価実験の結果，提案した t 検定に基づくトピックと局面の対応関係構築方法が，ベースライン手法に比べて高い推定性能を示すことを確認した．訓練データに単一ラベルを付与した場合と，複数ラベルを付与した場合で，JS Divergence によって確率分布の推定性能を評価した結果，特に単一ラベルという状況で階層的推定法は有意に良い推定ができることが明らかになった．

以上の結果から，ツイートのような短文に対して，より少ない訓練データでマル

チラベル分類をする場合や，確率分布推定をする場合に，提案した階層的推定法が有効であることを明らかにした．

Contents

1 Introduction **1**

2 Related works **7**

2.1 Information extraction from Twitter 7

2.2 Experience mining 9

2.3 Local information recommendation 10

2.4 Extracting information related to user's life 11

2.5 Hierarchical manner 12

2.6 Topic model 13

2.7 Multi-label classification 14

2.8 Summary 15

3 Hierarchical estimation framework **17**

3.1 Overview of HEF 17

3.2 Topic extraction using LDA 18

3.3 Relevance calculation 20

3.4 Relevance normalization 21

3.5 Association building 21

3.6 Aspect score calculation 23

4 Multi-label classification **25**

4.1 HEF extension for multi-label classification 25

4.1.1 Entropy feedback 25
4.1.2 Aspect estimation 27
4.1.3 Optimal number of topics 28
4.2 Experimental evaluations 29
4.2.1 Dataset and parameter settings 29
4.2.2 Evaluation metrics 31
4.2.3 Baseline methods 33
4.2.4 Experimental results 33
4.3 Discussions 54
4.3.1 Effectiveness of feedback entropy 54
4.3.2 Estimation performance of each method 55

5 Probability distribution inference 61
5.1 HEF extension for probability distribution inference 61
5.1.1 Optimal association building 61
5.1.2 Inference 62
5.2 Experimental evaluations 63
5.2.1 Dataset 63
5.2.2 Evaluation metrics 64
5.2.3 Baseline methods 66
5.2.4 Experimental results 67
5.3 Discussions 70

6 Discussions 77
6.1 Achievements in this dissertation 77
6.2 Associations with latent topics 79

7 Conclusion 81

Acknowledgements 83

Bibliography **85**

Publication List **99**

List of Tables

1.1 Aspects of real life 6

4.1 Number of correctly labeled aspects 31
4.2 Co-occurrence aspects ratio by examinee aspects 32
4.3 Compared to estimation performance of HEF using each threshold value 35
4.4 Precision of each method 46
4.5 Recall of each method 47
4.6 F-measure of each method 48
4.7 Number of labelings by each method 49
4.8 Relevance of high $\hat{Ra}$ topics built by HEF0 50
4.9 Relevance of high $\hat{Ra}$ topics built by HEF 51
4.10 High occurrence probability terms in each topic associated to many aspecs 52
4.11 High occurrence probability terms in each topic strongly associated to each aspect 58
4.12 Complete estimated aspects for tweet by HEF 59
4.13 Estimated extra aspects for tweet 59

5.1 Number and probability of labels by aspect 65
5.2 JSD scores in each number of topics in HEF 69
5.3 JSD by each association building method 69

5.4 Effectively inferred probability distributions of aspects by HEF . . . 75
5.5 High occurrence probability terms in highest relevance topic associated to Event . . . 75
5.6 Number of average labels for a tweet in each examinee . . . 76

List of Figures

3.1 Hierarchical estimation framework 18

3.2 Graphical model of LDA 20

3.3 Association examples 22

3.4 Aspect estimation method 23

4.1 JS_{sum}, Precision, Recall, and F-measure values of each number of topics 34

4.2 Converging state of feedback coefficients 35

4.3 Connectivity among topics and aspects 36

4.4 Precision, Recall, and F-measure of Appearance 38

4.5 Precision, Recall, and F-measure of Contact 38

4.6 Precision, Recall, and F-measure of Disaster 39

4.7 Precision, Recall, and F-measure of Eating 39

4.8 Precision, Recall, and F-measure of Event 40

4.9 Precision, Recall, and F-measure of Expense 40

4.10 Precision, Recall, and F-measure of Health 41

4.11 Precision, Recall, and F-measure of Hobby 41

4.12 Precision, Recall, and F-measure of Living 42

4.13 Precision, Recall, and F-measure of Locality 42

4.14 Precision, Recall, and F-measure of School 43

4.15 Precision, Recall, and F-measure of Traffic 43

4.16 Precision, Recall, and F-measure of Weather 44
4.17 Precision, Recall, and F-measure of Working 44
4.18 Precision, Recall, and F-measure of Other 45
4.19 Precision evaluated by varying amount of training data 53
4.20 Recall evaluated by varying amount of training data 53
4.21 F-measure evaluated by varying amount of training data 54

5.1 Relevance and t-test value distributions of Disaster 71
5.2 Relevance and t-test value distributions of Event 71
5.3 Relevance distributions of all aspects 72
5.4 t-test value distributions of all aspects 73
5.5 JS divergence 74
5.6 Euclidean distance 74
5.7 Probability distributions of aspects estimated by each method for **Table 5.4**'s tweet 76

Chapter 1

Introduction

Many information sharing services currently exist, such as community knowledge sharing sites, blogs, and microblogs. Twitter [2], which is one of the most popular social media services, had 280 million active users per month at the end of September 2013 [71]. Since Twitter only permits users to post short sentences up to 140 characters, users can easily post their experiences and opinions about daily events. Thus, Twitter posts are often both useful and timely because they typically discuss current events. For example, tweets about traffic jams or traffic accidents are quite valuable for users who will pass those places. Supermarket sales and bargain information are also helpful for neighborhood consumers. Such tweets, which are highly regional, up-to-date, and beneficial to others, are called real life tweets.

The Great East Japan Earthquake Disaster, which occurred in March of 2011 [1], is a actual example of the benefits of real life tweets. There was great amount of confusion in the stricken area immediately following the earthquake. There was a lack of food, suspension of water supply, and train service cancellations. At that time, useful tweets reported the location of water supplies and food distributions, as well as the service status of trains, demonstrating that such real life tweets helped the users in the devastated region [84].

Information is used in various aspects of life. Real life tweets can accommodate

such aspects. For example, tweets such as "Train is not coming!" are categorized in the "Traffic" aspect and will support users who want to ride the train. Posts such as "Today, bargain sale items are 50% off!" are categorized as "Expense" aspects and will support users who are going shopping. For presenting real life tweets based on user contexts, we classify them into 14 aspects. The 14 aspects shown in **Table 1.1** are assumed to be users life aspects that refer to the Yahoo directory [1], "local community" [2], and "life" [3]in the Japanese version of Wikipedia.

Users who provide particular aspect information sometimes can be found on Twitter. For example, in the Disaster aspect, Twitter officially collects users who posts lifeline information and create users lists in each prefecture and in each public institution in Japan. In the Weather aspect, the Japan weather association officially operates an account that posts the latest weather information. In the Traffic aspect, the Tokyo Transportation Bureau's account provides train information. However, these accounts generally post information that is already known to users. To detect daily local information, we also need to collect general user tweets because they function as sensors that observe daily events. In fact, several studies on event detection treat general users as social sensors to achieve their goal [62, 61, 51, 66].

On the other hand, users post various types of tweets. "Nod" and sympathetic phrases frequently appear on Twitter. For example, "Thank you" and "I see" often appear in posts. These posts do not directly support the real life situations of other users. We believe that users want a method of locating beneficial tweets on Twitter. These types of nods and sympathies simply impede the discovery of substantive tweets.

Users generally expect to get contents that reflect their particular interests. However, the above problems occurred in many applications and social media. For example, many spam e-mails are delivered daily. Some blogs offer no content of interest.

[1]http://business.yahoo.com

[2]https://ja.wikipedia.org/wiki/地域コミュニティ

[3]https://ja.wikipeda.org/wiki/生活

News articles mention various categories. These hinder not only the information access of users but also bury user interesting contents under too much information. These problems can be solved by estimating the labels of each bit of content so that users can rapidly access contents by labels of interests.

The objective of this research is to estimate the aspects of unknown tweets to provide real life tweets to users who expect particular aspects. The simplest idea to achieve this goal is to define terms with high relevance to each aspect by hand and calculate the aspect scores based on them. Although these approaches have achieved high accuracy in many studies, estimating real life tweets is hard because each aspect includes many keywords without completely enumerating the important terms for estimating aspects.

In this study, we address a supervised machine learning approach that trains a model by labeled data and estimates unknown data using the model. To estimate the aspects of unknown tweets, we follow these four challenges.

Challenge 1 We must estimate the aspects of unknown tweets from a few feature terms because 45, which is the average number of characters in tweets, is short compared with general documents [49].

Challenge 2 We must achieve the highest accuracy by the latest labeled tweets because the lives of people momentarily change and new terms might appear. Since annotating the aspects of many tweets is difficult because the aspects are hand-labeled, we need to estimate them by a small set of labeled tweets.

Challenge 3 Depending on the tweets, we have to estimate several aspects of a tweet. For example, the following tweet "A heavy snowstorm caused a traffic accident near the JFK airport," mentions a snowstorm and a traffic accident. Its main topic is the accident, but it also provides weather information. Therefore, we multi-label it as both Traffic and Weather.

Challenge 4 An approach that estimates several aspects of a tweet can clearly pro-

vide real life information for specific users. On the other hand, exhaustive-oriented users might expect broad information that includes the specific aspects. In other words, accuracy-oriented users might desire strictly selected real life information on specific aspects. When we visit sightseeing locations, we want information about them. The multi-label classification approach failed to achieve such tightly associated aspects with the same weight.

In typical supervised machine learning methods as document classification, naive Bayes classifier [17] and support vector machine (SVM) [15] are widely known. Both methods are extended to multi-labeling [33, 12] and labeled latent Dirichlet allocation (L-LDA) [58] also proposed to multi-label for documents. These methods address classification of relatively long documents and achieve high accuracy using enough training dataset. However, in cases of short document as tweet and a small set of labeled data, that is suggested that these methods cannot achieve adequate performance because feature terms of tweet is few.

In this study, we propose the hierarchical estimation framework (HEF) based on associations between topics and aspects to achieve above challenges. The fundamental idea of it is composed of both unsupervised and supervised machine learning techniques. In the first phase, it extracts topics from a sea of tweets using latent Dirichlet allocation (LDA). In the second phase, it calculates the relevance between topics and aspects using a small set of labeled tweets to build associations among them. HEF calculates aspect scores for unknown tweets using the association between topics and aspects based on the terms extracted from tweets.

Although typical machine learning methods directly calculate the likelihood of terms, HEF calculates relevance between topics and aspects using a small set of labeled tweets and build the associations based on relevance. HEF feature is to expand terms of tweet to topics, estimate contents in topics, and calculate aspect scores by associations between topics and aspects. Thus, by two phase estimation based on topics as middle-layer, even if a term do not appear training data, we can

calculate aspect scores using it. In other words, HEF stochastically expand terms of tweets using topics. Though, it tries to immoderately expand terms, estimation performance decrease. Therefore, we preliminarily calculate occurrence probability of terms in each topic by LDA using a large amount of tweets.

In multi-label classification, when the aspect scores exceed a threshold, the aspects are estimated for tweets. The HEF is introduced entropy feedback mechanisms in the second phase to overcome the problem of competitive associations among aspects. Based on these extensions, the associations between topics and aspects are refined and the estimation precisions are increased. We evaluate the Shannon entropy of each association between the aspects and topics and iteratively calculate the feedback coefficients by entropy to achieve optimal associations.

The aspect distribution is represented by the probability distribution in each tweet. Accurately inferring the probability distribution of the aspects means supporting either the strict or broad associations between tweets and aspects. As an inference approach of probability distribution, we naturally extend HEF by normalizing scores and propose an optimal association building method based on t-test, which is an efficient strategy to manage the relationship between topics and aspects. We assume that the training data are not given as the probability distributions of the aspects based on a training model of a typical classification method. Our challenge in this study is to train from labeled tweets and infer the probability distribution of the aspects of unknown tweets.

The organization of this dissertation is as follows. Chapter 2 describes related works. Chapter 3 proposes the fundamental of HEF. Chapter 4 extends HEF to multi-label classification by introducing entropy feedback and evaluates estimation performance using actual real life tweets. Chapter 5 explains probability distribution inference based on HEF and examines inference accuracy by comparing with other methods. Chapter 6 widely discusses the effectiveness of HEF, and conclude this study and briefly describe future works in Chapter 7.

Table 1.1: Aspects of real life

Aspect		typical terms
Appearance	(App.)	clothes, dress, wearing, fashion, uniforms, kimono, decoration, makeup, haircuts ...
Contact	(Con.)	appointments, meetings, invitations, family, friends, parties, drinking parties, get-togethers ...
Disasters	(Dis.)	flood, tornados, earthquakes, seismic ocean waves, power loss, hazards, secondary disasters ...
Eating	(Eat.)	cooking, dining out, eating, restaurants, recipes, ingredients ...
Events	(Eve.)	festivals, ceremonies, projects, schedules of events, conferences, special days, art shows ...
Expense	(Exp.)	shopping, orders, advertisements, discounts, bargains, markets, sales, purchases ...
Health	(Hea.)	colds, physical condition, aches and pains, hospital, health management method, medicine ...
Hobbies	(Hob.)	leisure-time, pastime, entertainment, hobbies, interest, games, music, television, movies ...
Living	(Liv.)	home, lodgings, furniture, cleaning, doing laundry, living, apartment, accommodation ...
Locality	(Loc.)	sightseeing, regionally specific, local information ...
School	(Sch.)	study, class, examinations, education, research, homework, coursework, lectures cancellation ...
Traffic	(Tra.)	trains, buses, airplanes, timetables, traffic information, clogs, roads, traffic jams, accidents ...
Weather	(Wea.)	weather forecasts, temperature, humidity, hail, rain, thunder, sky, air, wind, pollen ...
Work	(Wor.)	job hunting, part-timer, coursework, opening a store, closing a business, job, employment ...

Chapter 2

Related works

2.1 Information extraction from Twitter

The study of information extraction from Twitter is flourishing. Mathioudakis *et al.* [46] extracted burst keywords in automatically collected tweets and found trends that fluctuated in real time by creating groups using the co-occurrence of keywords. Zhao *et al.* [88] extracted tweets about information needs using a Support Vector Machine (SVM) to discover real world trends and events. Wang *et al.* [73] estimated user interests using posted tweets to discover effective users for tweet diffusion. Li *et al.* [42] proposed the extraction method for named entity posted to Twitter by unsupervised learning. They split a tweet to term chain and calculate its score based on mutual information. Based on the hypothesis that a named entity frequently co-occurs other named entity, they ranks named entities by co-occurrence frequency in tweets posted in periods. Rajadesingan *et al.* [57] detect the sarcasm in Twitter to help company's customer services. They introduce psychological studies and sentiment score of term into the modeling framework to discover the sarcasm. Bollen *et al.* [9] analyzed sentiment on Twitter based on a six-dimensional mood (tension, depression, anger, vigor, fatigue, and confusion) representation, and determined that on Twitter, it correlates with such real-worlds values as stock prices and coincides with cultural events. Wang

et al. [75] extracted bursty topics with high correlation by comparing burst patterns among different news streams for various viewpoints. Iwaki *et al.* [30] detected the beneficial articles posted in Twitter by calculating similarity between users and tweets posted in previous. Sriram *et al.* [64] classified tweets into five categories such as news, events, opinions, deals, and private messages by preliminarily extracting features from author's profile and text. Li *et al.* [41] detected the burst intervals whose combinations of words rapidly increased in Twitter. They similarly detected the events for newly obtained document streams, calculated the similarity between these and old events, and tracked them. Xie *et al.* [81] also proposed a topic tracking method called *TopicSketch* to achieve the same purpose with low calculation costs. Their method detects the bursty topics by concurrently observing all Twitter streams and the documents of each term and each term's pair. In this paper, we estimate real life aspects of unknown tweets.

Several studies are focusing on user interests at Twitter. Hannon *et al.* [20] proposed the users recommendation method using their past tweets and followee/follower. They calculated the weight for each feature by machine learning and achieved the high recommendation performance. Michelson and MacsKassy [47] discovered user's topics on Twitter by categorizing the entities in the tweets and developing user profiles by adopting categorization results. Wu *et al.* [80] automatically generated personalized tags to label Twitter's user interests. They extracted keywords from Twitter messages and calculated TF-IDF and TextRank [48] scores for them. Yamaguchi *et al.* [82] proposed a user tagging method using Twitter lists to discover user topics. Based on their observations, they assumed that the users included on identical lists probably posted on the same topic. From experimental evaluations with two datasets, their method effectively acted as a user tagging method. Cha *et al.* [10] analyzed user features with influence by comparing the number of followers, followees, and replies. Those users with maximum influence wield critical power on various topics. They also clarified that influence cannot be obtained by only posting on a single topic. The

objective of this research is to infer the aspects of tweet without reference to users.

2.2 Experience mining

Real life tweets consist of both the experiences and the knowledge of users. Several studies on experience mining have extracted experiences from documents. Kurashima *et al.* [39] divided human experience into five areas: time, space, action, object, and feeling. Inui *et al.* [27] indexed personal experience information from the viewpoints of time, polarity, and speaker modality. This information is indexed as topic object, experiencer, event expression, event type, and factuality. Ikeda *et al.* [26] and Takano *et al.* [67] defined the combination rules of word class to extract sentences about user's experience. They extracted sentences as experience information when they include rules. Hattori and Nadamoto [21] extracted important and unique information related to social media as tip information and comments including user experiences by using common important words. The *et al.* [69] automatically extracted all the basic attributes, such as actor, action, object, time, and location from weblogs using conditional random fields and self-supervised learning. Nishihara *et al.* [53] proposed a support system for obtaining personal experience from blogs using images. They extracted terms that represent events including place, object, and action, and provided such images as term objects to users. Their system showed high effectivity in obtaining personal experiences.

These mining methods are effective for relatively long documents such as blogs. But they are inappropriate for Twitter posts, which consist of many short sentences. In addition, experience mining is much more difficult because subjects and objects are often omitted in Twitter sentences.

Arimitsu *et al.* [6] proposed a retrieval method for experiences that users fragmentally posted in Twitter. They argued that one experience is consisted of several behavior transitions, and their method searches for articles including specific ex-

periences by inputting the keyword chains into their system. However, they don't automatically generate keyword chains related to experience.

2.3 Local information recommendation

Real life tweets often include the fresh local information. Many studies related to its extraction exist. Kurashima *et al.* [38] recommended locations from users' location log data. They proposed a probabilistic behavior model based on both user interests and the activity area that hosts user homes, offices, and so on. Ma and Tanaka [43] proposed a regional information retrieval method from web pages utilizing the current location data and keywords of users. Regional information was ranked based on the notion of localness degree. Lee *et al.* [40] proposed an urban regional characteristics extraction method using tweets whose geo-locations were tagged. Their method extracted the feature behavior patterns in the region by analyzing crowd behavior in tweets.

As a feature of location information, regional and establishment names that only exist in only specific regions are used. The study of the extraction of these feature terms in locations is flourishing. Oku *et al.* [54] defined high regional terms whose occurrence frequency in the target regions is relatively high compared with other regions and proposed a regional score calculation method for terms by introducing inverse document frequency (IDF) and term co-occurrence frequency with municipality names. Cheng *et al.* [13] detected the terms for specific regions from geo-tagged tweets and obtained the features to determine regional attributions by their proposed methods. Doumae and Seki [18] automatically selected feature topics for specific regions to estimate the user's life area by applying the Dirichlet Process with Mixed Random Measures (DP-MRM) [34] which is a semi-supervised topic model. Yamaguchi *et al.* [83] detected the feature terms in spatiotemporal areas to infer user locations. Their method assumed social media contents that are generated in real

time, and the terms are continuously updated by newly arriving contents. Arakawa *et al.* [4] extracted keywords with high dependency for locations by splitting them into 100-km grids and calculated the content percentage of the geo-tagged tweets in each grid.

This research estimates not just for the Locality but for 14 real life aspects.

2.4 Extracting information related to user's life

Many studies have extracted beneficial information for the lives of users. Extracting traffic information from social media has been particularly widely studied. Sakaki *et al.* [61] extracted real-time driving information from social media to provide current traffic situations to users. Their developed system incorporated geographically related terms into geographical coordinates. Nagano *et al.* [51] developed a system for detecting railway information from train schedules from Twitter. Their system collects the latest tweets posted within three minutes, including railway names. When the number of tweets, which satisfied their three defined rules, exceeded their threshold, they judged that a target railway is late. Tsuchiya *et al.* [70] classified train problems into stoppages, partial suspensions, and other troubles using a SVM that is trained by tweets including railway names.

Ishino *et al.* [29] proposed a transportation route extraction method during disaster by tagging such terms as departures, destinations, and transportation devices. Sakaki *et al.* [62] assumed that Twitter users act as social sensors that identify both earthquakes and typhoons in real time in the real world. They estimated the occurrence locations and the period of these events. Kawaguchi *et al.* [32] proposed an information collecting system during disasters using Twitter. Their system extracts tweets including location names related to user attributions and situations.

Aramaki *et al.* [5] predicted influenza epidemics using Twitter. They extracted tweets related to influenza based on an SVM modeled by tweets that literally mention

influenza patients. Takahashi and Noda [66] developed a pollen visualization system on a Japan map using Twitter. The tweets related to pollen are collected by SVM.

Nakajima *et al.* [52] recommended the travel routes by collecting tweets posted on popular sightseeing spots and classifying them into three categories (eating, landscape, and activity) using feature terms and parts of speech. Ishino *et al.* [28] extracted both information of souvenir and tourist spots as travel information from travel blog entries. Moreover, they built a collection of travel information links by extracting hyperlinks from travel blog entries.

Although these studies extract beneficial information in particular life aspects, our research concurrently estimates several aspects of unknown tweets based on multi-label classification and probability distribution.

2.5 Hierarchical manner

Hierarchical frameworks have been adopted in many studies to achieve various tasks. Chan *et al.* [11] proposed term selection and weighting methods using hierarchical category classification to achieve question retrieval and ranking in community question answers. Ren *et al.* [59] proposed a hierarchical multi-label classification method which considered three core factors: short document expansion, time-aware topic tracking, and chunk-based structural learning. Zhu *et al.* [89] generated topics from a social media corpus and constructed a topic hierarchy in the information of each user need in such noun phrases as “iPhone 5” and “Facebook Inc.”. They achieved optimal topic hierarchy by calculating its likelihood based on the weight of the edges between subtopics. Wang *et al.* [74] iteratively split the topic set into subtopics to build a hierarchy of topics. Their constructed hierarchy is integrated by a ranked list of mixed length phrases. Hu *et al.* [25] achieved intent-aware search result diversification for information retrieval systems by hierarchically representing user intents and proposed two hierarchical diversification models based on the novelty and popularity

of each topic.

Our hierarchical estimation framework (HEF) consists of tweets, topics, and aspects. HEF extracts topics from a sea of tweets using topic models and associates topics to aspects. One of the HEF's features is to build the appropriate associations between many topics and aspects for estimating the aspects of unknown tweets.

2.6 Topic model

Topic model studies widely use LDA [8], which is a latent topic extracting method that was devised for a probability topic model. LDA supposes that a document is a mixture distribution of plural topics. Each topic is expressed by the probability distribution of the terms. Zhao *et al.* [87] proposed a model called Twitter-LDA, based on the hypothesis that one tweet expresses one slice of a topic's content. They classified tweets by topics and extracted keywords to express their contents. Diao *et al.* [16] detected bursty topics using Time-User-LDA, which is an extension of LDA. They evaluated the accuracy of topic detection among three LDA models and clarified that Time-User-LDA detects with the highest accuracy. Ma *et al.* [44] automatically annotated hashtags to tweets. Their PLSA-style models include user, time, and tweet content factors and achieved higher precision than other methods. Hong and Davison [23] evaluated how the restricted length of tweets limits the potential of traditional topic models and showed that training a topic model on aggregated messages significantly enhanced the experiment performance.

Topic model is applying to many studies. Kimura and Miyamori [35] classified the relationship between hashtags into four classes such as similarity, conflict, relevance, and irrelevance by estimating topic distribution of hashtags using LDA. Koike *et al.* [36] extracted the bursty topics with a correlation between news streams and Twitter by applying dynamic topic model (DTM) [7], which analyzes topic distribution transitions of each document on time-series. Weng *et al.* [78] estimated user topics

using LDA and detected the users who exert great influence on Twitter. They built a network for each topic based on follows and followers and calculated each user's score in each network using *TwitterRank* which extended PageRank [55]. Zhang *et al.* [86] recommended bands to music lovers using LDA by calculating the degree of artist similarity based on generated topics. Users received recommendations about artists in whom they might be interested. Pennacchiotti and Popescu [56] classified the user's political orientation based on user-centered attributes (profile, vocabulary, behavior, and sociality) using LDA in feature selection of SVM. Riedl *et al.* [60] found the change-points of topics using LDA by calculating the similarity between sentences that express the vectors of topic frequency. In this paper, we build associations between aspects and topics generated by LDA.

2.7 Multi-label classification

Multi-label classification studies are widely known methods based on SVM, naive Bayes classifiers, and LDA. SVM, which is one identification method that performs supervised learning, has high generalizing capability and classification performance [15]. Chang *et al.* [12] developed a SVM library called LIBSVM, which achieves multi-label classification by building models by combining several labels.

A naive Bayes classifier assumes that the term occurrence in a document is independent, and label probabilities are calculated from these terms using Bayes rules. It estimates labels with the highest probability for a document [17]. Wei *et al.* [76] proposed multi-label classification based on naive Bayes classifiers and estimated several labels with the probability that exceeds the average score calculated by all the label probabilities. Hong *et al.* [22] propose a multi-label classification based on probabilistic approach using conditional tree-structured Bayesian networks. They build the hypothesis that each label is depending on others and showed higher performance than other methods.

Ramage *et al.* [58] suggested a model called Labeled LDA (L-LDA) that expanded LDA to supervised learning. To extract latent topics, it assumes the labels to be the contents of documents. L-LDA can extract a one-to-one correspondence between LDA's latent topics and document labels.

Kase and Miura [31] estimated the new labels for existing news-corpus. They calculate occurrence probability of each feature in each class from multi-label dataset and estimate additional label with high probability using EM algorithm based on multinomial mixture model.

These methods show high estimation performance of such long documents as blogs and newspapers using sufficient training data. However, tweets consist of fewer terms because their length averages 45 characters [49]. Moreover, as training data, fresh tweets are preferred because they are easily influenced by the real world. In these conditions, typical multi-label classification methods fail to produce adequate performance to estimate several aspects of unknown tweets [85].

2.8 Summary

This dissertation studies the life aspect inferences of unknown tweets by a hierarchical estimation framework (HEF), which consists of an unsupervised topic model and associations built by supervised learning. One of the most characteristic points of this dissertation is how we construct effective associations between topics and aspects for estimating the aspects of unknown tweets in HEF. We describe a construction method of refined associations by iteratively calculating entropy based on the current associations in Chapter 4. We also explain the extraction method of the optimal topic set for each aspect by a t-test in Chapter 5. As an inference approach, HEF deals with multi-label classification and the probability distribution inference of the aspects. Users can get real life tweets with freshness and high regionality because these two approaches can be applied to various user orientations. Although tweets

are shorter than general documents, HEF expands the terms using topics generated by LDA and uses them for estimations. Therefore, our proposed scheme will achieve higher completeness than previous approaches.

Chapter 3

Hierarchical estimation framework

3.1 Overview of HEF

The overview of hierarchical estimation framework is shown in **Fig. 3.1**. In the first phase of HEF, a large numbers of topics are extracted from a sea of tweets using LDA. In its second phase, associations between topics and aspects are constructed using a small set of labeled tweets. We calculated the aspect scores for unknown tweets using the associations based on the terms extracted from them.

Typical supervised machine learning methods directly calculate the term likelihood from labeled training data. The terms in unknown tweets, which do not appear in the training data, cannot play a effective role in the estimation of conventional methods. In contrast, HEF is composed of a triple hierarchy: Tweet-Topic-Aspect. The terms in a tweet are expanded using co-occurrence terms in appropriate topics. From these reasons, we clarified that HEF can estimate several appropriate aspects from a small set and the short sentences of labeled data: i.e., tweets.

The organization of this section is as follow. Section 3.2 compactly explains topic extraction using LDA. Section 3.3 calculates relevance between topics and aspects. Section 3.4 describes necessity of relevance normalization. Section 3.5 build associations between topics and aspects based on relevance. Section 3.6 explains calculation

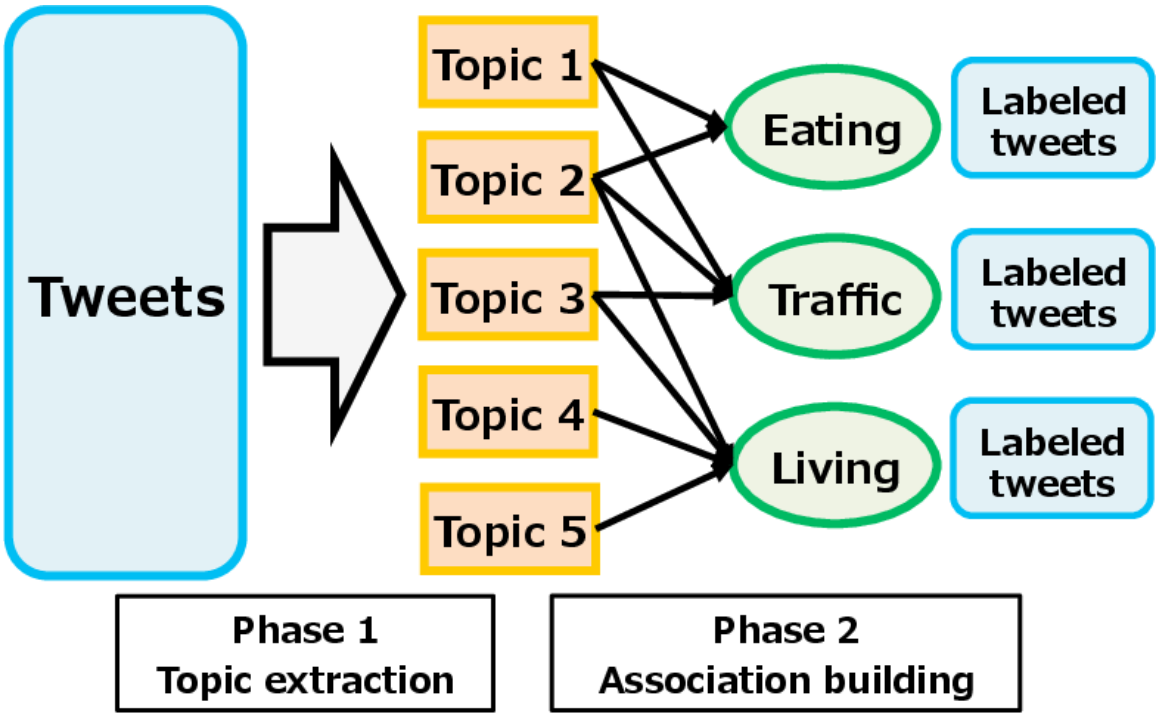

Figure 3.1: Hierarchical estimation framework

method of aspect scores of tweet.

3.2 Topic extraction using LDA

LDA is a latent topic extracting method using a probability topic model devised by Blei. A graphical model of it showed in **Fig. 3.2**. LDA supposes a document to be a mixture distribution of plural topics. Each topic is expressed by the probability distribution of the terms.

A document-term model consisting of n documents and m terms is expressed by a $n \times m$ matrix. In LDA, r latent topics are generated document-term models represented by $n \times r$ and $r \times m$ matrices. The fundamental idea of LDA is that documents are expressed in a mixture topic distribution and topics are expressed by the probabilistic distribution of terms. LDA applies a Dirichlet prior on the multinomial distribution over the topics for the documents. The LDA proceeds in the following

steps:

1. For all documents d sample $\theta_d \sim \text{Dirichlet}(\alpha)$
2. For all topics t sample $\phi_t \sim \text{Dirichlet}(\beta)$
3. For each N_d term w_i in document d:
   (a) Sample topic $z_i \sim \text{Multinomial}(\theta_d)$
   (b) Sample topic $w_i \sim \text{Multinomial}(\phi_z)$

where α and β are the hyper parameters for the Dirichlet prior. To generate an LDA model, we must estimate topic collection Z. We used a collapsed Gibbs sampling method[19]. Probability $P(z_i = k|Z_{-i}, W)$, in which the nth term in document d belongs to topic $z_i = k$, can be calculated as follows:

$$P(z_i = k|Z_{-i}, W) = \frac{N^d_{k-i} + \alpha}{N^d_{-i} + T\alpha} \cdot \frac{N^d_{k-i} + \beta}{N^d_{k-i} + W\beta} \tag{3.1}$$

, where i means the nth term in document d. N^d_{k-i} denotes the number of assignments of topic k in document d without term i, N^d_{-i} counts the terms in document d without term i, N^v_{k-i} represents the frequency of term v in topic k without term i, and N_{k-i} denotes the number of terms in topic k without term i. T and W are the number of topics and the vocabulary.

Term-topic distribution ϕ and topic-document distribution θ are estimated from topic collection Z calculated by a collapsed Gibbs sampling. Probability $\hat{\phi^w_k}$, whose term t is generated from topic k, and $\hat{\theta^k_d}$, whose topic k is generated from document d, are estimated as follows:

$$\hat{\theta^k_d} = \frac{N^d_k + \alpha}{N^d + T\alpha} \quad , \quad \hat{\phi^w_k} = \frac{N^v_k + \beta}{N_k + W\beta}. \tag{3.2}$$

Topic k is expressed as the occurrence probability of terms. All terms have probability in all topics.

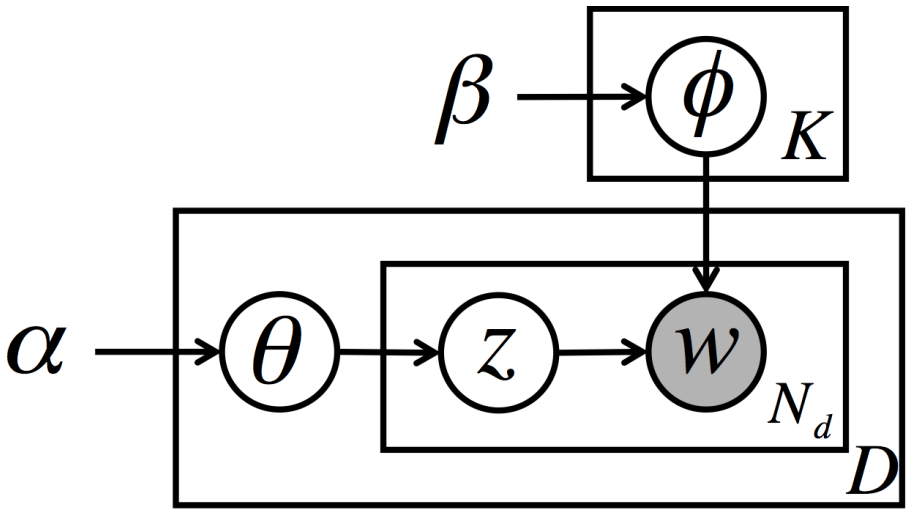

Figure 3.2: Graphical model of LDA

3.3 Relevance calculation

For building associations, relevances between topics and aspects are calculated using a small set of labeled tweets. A set of extracted terms from tweets is W. Relevance $R(a,t)$ between topics t and aspects a is calculated as follows:

$$R(a,t) = \sum_{w \in W} p(a,w)^{\alpha} * p(t,w)^{\beta}, \tag{3.3}$$

where $p(t,w)$ denotes the occurrence probability of term w in topic t preliminarily calculated by LDA. $p(a,w)$ denotes the occurrence probability of term w in aspect a calculated by a small set of labeled tweets and is calculated as follows:

$$p(a,w) = \frac{n_{w,a}}{\sum_{w' \in W} n_{w',a}}, \tag{3.4}$$

where $n_{w,a}$ denotes the occurrence number of term w in tweets where aspect a is labeled. Note this equation only calculates the relevance between topics and aspects using the occurrence probability. α and β, which are feedback coefficients to control the extent of occurrence probability, are calculated in Section 4.1.1.

3.4 Relevance normalization

To build the optimal associations between topics and aspects, we focus on the relationship between topics and aspects. The relevance $R(a,t)$ calculated by Eq. (3.3) is greatly different by feature in each topic t and aspect a. In an aspect side, when number of labeled tweets in particular aspect a is much compared with other aspects, relevance $R(a,t)$ is large value compared with other aspects because Eq. (3.3) calculates summation values. The importance $\hat{Ra}(a,t)$ of topic t in each aspect a is obtained by normalizing for all topics T and is calculated as follows:

$$\hat{Ra}(a,t) = \frac{R(a,t)}{\sum_{t' \in T} R(a,t')}. \tag{3.5}$$

Next, we consider a topic side. For example, topic, which are aggregated by location names extracted by LDA with high occurrence probability, are connected with high relevance to many aspects because real life tweets often contain location names. Similarly, topics including stop-words [65] will be associated with strong relevance to many aspects. This problem can be solved by normalizing for all aspects A. The normalized relevance $\hat{Rt}(a,t)$ is calculated as follows:

$$\hat{Rt}(a,t) = \frac{R(a,t)}{\sum_{a' \in A} R(a',t)}. \tag{3.6}$$

3.5 Association building

We make associations between topics and aspects. Here, depending on the aspects, note that the associations with topics are different. For example, the Eating aspect may be supported by fewer topics with high probabilities, and the Living aspect may be supported by many topics with mid-level probabilities (**Fig. 3.3**). We must construct various associations of each aspect because the optimal topic set is different for each aspect.

Therefore, we make an association between topics and aspects when $\hat{Ra}(a,t)$ ex-

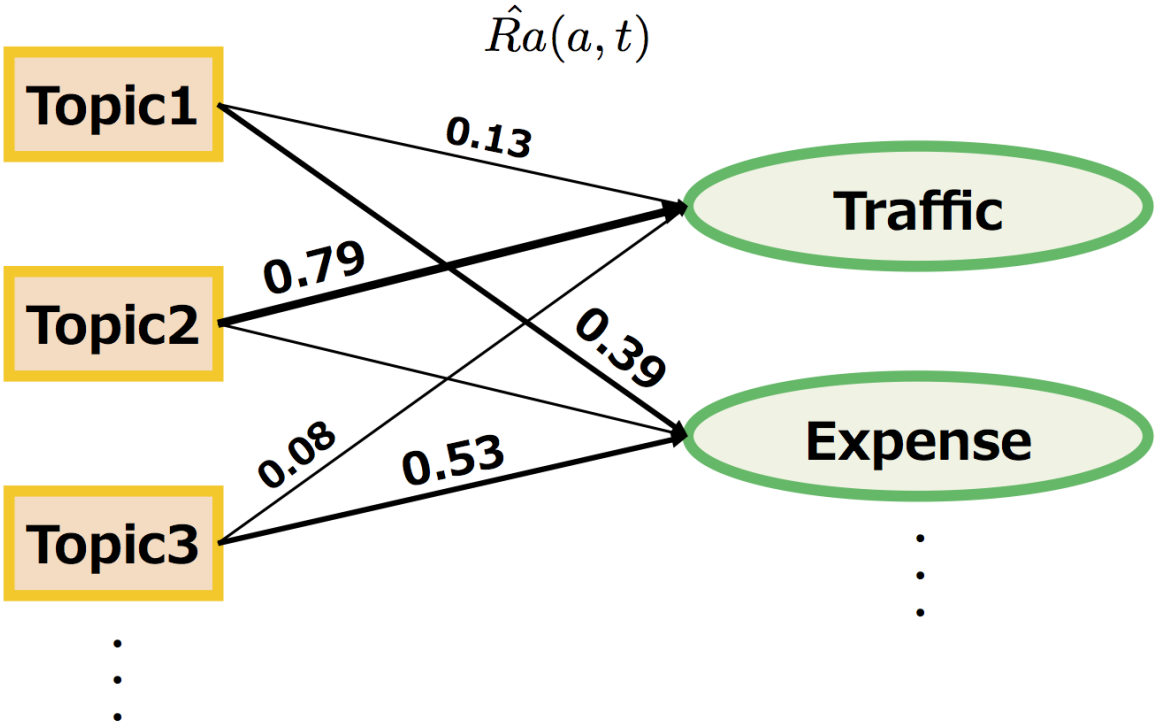

Figure 3.3: Association examples

ceeds a calculated threshold in each aspect a. Topic set T_a of aspect a is shown as follows:

$$T_a = \{t|\hat{Ra}(a,t) > \max_{t \in T} \hat{Ra}(a,t) - \sigma(\hat{Ra}(a,T)) * d\}, \tag{3.7}$$

where $\sigma(\hat{Ra}(a,T))$ denotes the standard deviation in $\hat{Ra}(a,t)$ for all topics. $\sigma(\hat{Ra}(a,T))$ play a role of feature value to represent relevance distribution of each aspect a. When $\sigma(\hat{Ra}(a,T))$ is a high value compared with other aspects, aspect a is associated to specific topics with high relevance. Thus, aspect a is supported by fewer topics.

According to increase the parameter d, aspects are associated to more topics. The optimal value of d is caused when associations between topics and aspects achieve the maximum estimation performance.

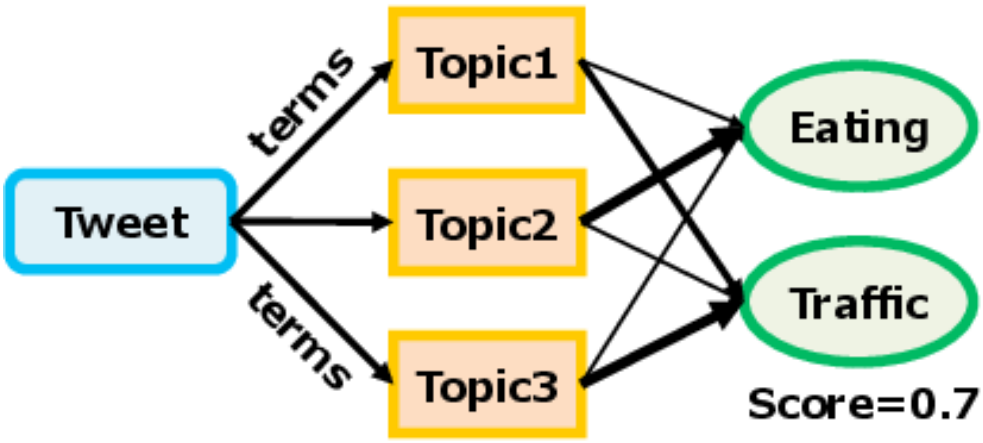

Figure 3.4: Aspect estimation method

3.6 Aspect score calculation

To estimate the aspects of unknown tweets, we use the associations between topics and aspects. The estimation flow using the associations is shown in **Fig. 3.4**. First, nouns, verbs, and adjectives are extracted from tweets. Second, the occurrence probabilities of all the terms are calculated for each topic. Then, the aspect score is calculated based on the tweet's probabilities and associations. Aspect scores $S(tw, a)$ between tweets tw and aspects a are calculated as follows:

$$S(tw, a) = \sum_{t \in T_a} \sum_{w \in W_{tw}} p(t, w)^{\beta} * \hat{Ra}(a, t) * \hat{Rt}(a, t), \tag{3.8}$$

where W_{tw} denotes a set of terms extracted from unknown tweet tw and $p(t, w)$ denotes the occurrence probability of terms w in topic t. β denotes the feedback coefficient calculated by Eq. (4.3).

$\hat{Ra}(a, t)$ gives high relevance to important topics for aspects. However, several aspects might strongly associate with the same topics. For example, topics in which verbs have a high rank of occurrence probability are given high relevance from many aspects because verbs often appear in many aspects. We believe that these topics decrease the estimation precision of aspects. $\hat{Rt}(a, t)$ also gives high relevance to the

characteristic topics of aspects, and low relevance to topics that share several aspects. Here, we must consider the properties of real life aspects with examples. For example, flood and heavy rain often appear in the same sentence because floods are generally caused by heavy rain; they are aggregated in the same topic by LDA. From **Table 1.1**, because flood and heavy rain are respectively included in Disaster and Weather aspects, both should share flood and heavy rain topics. However, $\hat{Rt}(a,t)$ gives low relevance to Disaster and Weather aspects.

To consider the relevance of both $\hat{Ra}(a,t)$ and $\hat{Rt}(a,t)$, we multiply both the relevances of the score calculation with Eq. (3.8).

Chapter 4

Multi-label classification

4.1 HEF extension for multi-label classification

4.1.1 Entropy feedback

Although HEF consists of the flows that achieved high estimation performance in average values, we clarified the aspects with low estimation performance [85] and the following two problems:

Problem 1 In relevance $\hat{Ra}(a,t)$ that was calculated by both Eq. (3.5), topics appear that are strongly associated to many aspects. Therefore, the topics are competitive among particular aspects. For example, the Disaster, Event, Locality, and Traffic aspects were associated with similar topics, which had such regional names as "Kyoto" and "Shijo". Tweets indicating these aspects probably appear in the regional names in a sentence because real life tweets mention the real world. This problem caused incorrect estimations and lowered estimation precision.

Problem 2 The aspects represented by some topics cannot build suitable associations because the relevance to the topics explained in problem 1 has high value. For example, in the case of the Disaster aspect that is appropriately repre-

sented by such topics as earthquakes and typhoons, the verb topics in problem 1 have higher relevance than these topics. Since tweet scores are calculated by un-feature terms, the estimation performance fell.

To solve problems 1 and 2, we can exclude terms by preliminarily defined stop-words. However, it is difficult for Twitter, which is a fast-changing medium. Moreover, since LDA generates topics including stop-words with high occurrence probability [65], we can exclude these topics with many stop-words. Although these topics do not need many aspects, we believe that they are important for particular topics. Therefore, we cannot completely exclude them.

LDA primarily provides high occurrence probability for high frequency terms in the dataset. Regional location terms have higher occurrence probability because they often appear in tweets. Therefore, to accurately calculate the relevance between topics and aspects, HEF has two kinds of parameters, α and β (Eq. (3.3)). In this paper, we propose a feedback method using Shannon entropy [63] to determine these parameters.

Entropy can evaluate the untidiness of probability distribution. $\hat{Ra}(a,\cdot)$ and $\hat{Rt}(\cdot,t)$ express the probability distribution in each aspect a and topic t. The entropies of both $H(a)$ and $H(t)$ are defined as follows:

$$H(a) = -\sum_{t\in T} \hat{Ra}(a,t) * \log_2 \hat{Ra}(a,t), \tag{4.1}$$

$$H(t) = -\sum_{a\in A} \hat{Rt}(a,t) * \log_2 \hat{Rt}(a,t). \tag{4.2}$$

Here, we must consider the association balance from some topics to an aspect. For example, as mentioned above, if such special terms as location names have high occurrence probability, the relevance is greatly high and entropy is low. Such association creates an unbalance for all the aspects. Hence, to control the occurrence probability of the terms, we calculate the feedback coefficients of both α and β on the basis of minimum entropy. α and β are calculated as follows:

$$\begin{aligned} \alpha &= \frac{1}{|A|}\sum_{a\in A}\frac{MA}{H(a)} \quad , \quad MA = \min_{x\in A} H(x), \\ \beta &= \frac{1}{|T|}\sum_{t\in T}\frac{MT}{H(t)} \quad , \quad MT = \min_{x\in T} H(x), \end{aligned} \tag{4.3}$$

where $|A|$ and $|T|$ denote the number of aspects and topics.

If the entropy difference of all the aspects and topics is increased, α and β are decreased. When both feedback coefficients are introduced to Eq. (3.3) within 1.0, the difference of the occurrence probability in the topics or the aspects is reduced; α and β lower the effectivity of the terms with especially high occurrence probability, such as place names. As a result, the entropy difference of every aspect and topic decrease, and the association balance of every aspect is preserved. Suitable associations between aspects and topics are built when α and β converge.

HEF is iteratively calculated in the order of Eqs. (3.3), (3.5), (3.6) (4.1), and (4.3). When α and β sufficiently converge compared to previous iteration values, HEF builds associations between topics and aspects by Eq. (3.7).

4.1.2 Aspect estimation

Aspects with high scores should be estimated for tweets. We estimate the top K aspects that are flexibly decided. In HEF, each aspect a score $S(tw, a)$ is normalized using score average $\mu(S(tw, A))$ and standard deviation $\sigma(S(tw, A))$. If the normalized aspect score exceeds each aspect's threshold $r(a)$, aspects are more likely to be estimated for the tweet. Some aspects A_{tw} for unknown tweet tw are estimated as follows:

$$A_{tw} = \left\{ a \,\middle|\, \frac{S(tw, a) - \mu(S(tw, A))}{\sigma(S(tw, A))} > r(a) \right\}. \tag{4.4}$$

Depending on the aspects, the estimation probabilities of the labels are intrinsically different. HEF decides threshold $r(a)$ in each aspect a from the number of labels $L(a)$ in the training data. Each aspect threshold $r(a)$ is calculated as follows:

$$r(a) = \frac{\mu(L(A)) - L(a)}{\sigma(L(A))}, \tag{4.5}$$

where $\mu(L(A))$ and $\sigma(L(A))$ denote both labels average and standard deviations of labels. This equation subtracts the number of each labeling aspect $L(a)$ from average value $\mu(L(A))$; the threshold is high when the number of labelings is less, and it is low when the number of labelings is great.

4.1.3 Optimal number of topics

LDA needs the number of topics as a parameter, which is important for our method because associations between topics and aspects are based on relevance. If the number of topics changes, the number associated with the aspects also changes.

Teh et al. [68] proposed the HDP(Hierarchical Dirichlet Process)-LDA to automatically optimize the number of topics for LDA by stratifying parameters. Although this method can decide the optimal number in LDA model, it is not necessarily optimal for our proposed method built by the association between topics and aspects.

To select the best number of topics in LDA for our proposed method, we used the JS Divergence [50] between each aspect and applied it to calculate the similarity between one aspect and others. When the JS Divergence is high, the probability distribution among aspects is much different. When it is 0, the probability distribution is identical. In this case, the maximum value of the JS Divergence sum indicates the optimal aspect set. Probability distributions use the $\hat{Ra}(a,t)$ of the aspects and the topics matrix. JS Divergence sum JS_{sum} is calculated as follows:

$$\begin{aligned} JS_{sum} &= \sum_{(\forall p, \forall q) \in A} D_{JS}(\hat{Ra}(p,\cdot), \hat{Ra}(q,\cdot)), \\ D_{JS}(x,y) &= \frac{1}{2}\left(\sum_{t \in T} x(t) \log \frac{x(t)}{z(t)} + \sum_{t \in T} y(t) \log \frac{y(t)}{z(t)}\right), \end{aligned} \tag{4.6}$$

where $z(t)$ denotes the average of $x(t)$ and $y(t)$.

4.2 Experimental evaluations

To clarify the effectiveness of our HEF which introduced feedback entropy method, we evaluated the precision, recall, and the F-measure values of the estimated aspects. As baseline methods, we used L-LDA, SVM, and NBML. By analyzing the associations between topics and aspects, we clarified the aspects for which the entropy feedback method was effective.

4.2.1 Dataset and parameter settings

Collecting many regional tweets

Our method requires many tweet datasets for generating topics using LDA. We collected 2,390,553 tweets posted from April 15, 2012 to August 14, 2012 using the Search API [3] on Twitter, each of which has "Kyoto" as the Japanese location information.

Real life tweets

To construct associations between the extracted topics and aspects, we prepared a small set of 1,500 labeled tweets, each of which has "Kyoto" as the Japanese location information. We used three examinees: examinee E1 is the first author, and E2 and E3 are university students living in Tsukuba City. During the labeling process, the examinees freely consulted **Table 1.1** and viewed the example tweets in each aspect and why they were classified as such. They selected the most suitable aspect for each tweet as the first aspect and the next two most suitable aspects as the second and third aspects. If no suitable aspect remained, they selected "other" to identify it as a non-real life tweet. Aspects that do not correspond to any candidate are listed fourth.

We evaluated the κ coefficients among the first candidates of the examinees [14]. When the κ coefficient is high, the classification agreement rate among the examinees is also high. The κ coefficient for examinees E1 and E2 was 0.687; it was 0.595 for examinees E1 and E3 and 0.576 for examinees E2 and E3. The average was 0.619,

which is a *substantial* match rate.

To appropriately give aspects to each tweet, we used the results from the labeling of all three examinees. Correct aspects AC_{tw} of each tweet tw are shown as follows:

$$AC_{tw} = \{a | Uscore(tw, a) \leq 10\},$$
$$Uscore(tw, a) = \sum_{u \in U} candidate(tw, a, u), \tag{4.7}$$

where U denotes all the examinees. $candidate(tw, a, u)$ is a candidate number: the 1st, 2nd, 3rd, and 4th rankings of aspects a labeled by examinee u for tweet tw. Hence, maximum $Uscore(tw, a)$ is 12 with three examinees when all $candidate(tw, a, u) = 4$. Minimum $Uscore(tw, a)$ is three when all $candidate(tw, a, u) = 1$.

For this determination, the number of labeling aspects of 1,500 tweets is shown in **Table 4.1**. The number of labels in the Appearance aspect is 181. The minimum number of labels is 86 in the Disaster aspect. The number of all labels for 1,500 tweets is 5,092, and the per tweet average of the labels was 3.39.

Next, we examined the co-occurrence aspects labeled by the examinees. The probability of co-occurring with other aspects is shown in **Table 4.2**. There are three rank columns, each of which shows an aspect having a top, second, and third probability in each aspect. The Appearance aspect co-occurs with the Expense aspect at 0.365 probability. The co-occurrence probability between Disaster and Weather, and between Traffic and Locality exceeds 0.5. These two aspects are concurrently mentioned in a tweet. The Locality aspect appeared with the high co-occurrence in other aspects: Disaster rank 2, Eating rank 3, Event rank 1, and so on.

Parameter settings

LDA requires hyperparameters. Based on related works [19], we set α to $\frac{50}{|T|}$ and β to 0.1. $|T|$ denotes the number of topics, chosen based on JS_{sum} from among 50, 100, 200, 500, and 1,000 topics in the **Section 4.2.4**. The iterative calculation count in LDA is 100 times in every case.

Table 4.1: Number of correctly labeled aspects

Aspect	Label	d	$\|T_a\|$
Appearance	181	17	451
Contact	379	3	16
Disaster	86	5	5
Eating	287	8	35
Event	311	10	159
Expense	435	13	500
Health	177	10	104
Hobby	348	17	500
Living	213	16	500
Locality	432	10	80
School	195	6	11
Traffic	169	10	5
Weather	226	5	2
Working	262	12	500
Other	1,391	1	12
Total	5,092		

4.2.2 Evaluation metrics

To discuss the effectiveness of our method, we evaluated the precision, recall, and F-measure values [45] in each aspect and these are calculated as follows:

$$\text{Precision}(a) = \frac{|\{tw \in D : a \in AC_{tw} \wedge a \in A_{tw}\}|}{|\{tw \in D : a \in AC_{tw}\}|}, \tag{4.8}$$

$$\text{Recall}(a) = \frac{|\{tw \in D : a \in AC_{tw} \wedge a \in A_{tw}\}|}{|\{tw \in D : a \in A_{tw}\}|}, \tag{4.9}$$

Table 4.2: Co-occurrence aspects ratio by examinee aspects

	Co-occurrence Rank 1		Co-occurrence Rank 2		Co-occurrence Rank 3	
	Aspect	Probability	Aspect	Probability	Aspect	Probability
Appearance	Exp.	0.365	Hob.	0.232	Liv.	0.155
Contact	Eve.	0.335	Hob.	0.303	Exp.	0.219
Disaster	Wea.	0.523	Loc.	0.430	Tra.	0.174
Eating	Exp.	0.446	Con.	0.279	Loc.	0.247
Event	Loc.	0.460	Con.	0.408	Hob.	0.283
Expense	Eat.	0.294	Loc.	0.269	Hob.	0.221
Health	Eat.	0.249	Liv.	0.209	Wea.	0.192
Hobby	Con.	0.330	Exp.	0.276	Eve.	0.253
Living	Exp.	0.197	Wea.	0.183	Hob.	0.183
Locality	Eve.	0.331	Exp.	0.271	Tra.	0.245
School	Con.	0.323	Wor.	0.251	Eve.	0.190
Traffic	Loc.	0.627	Eve.	0.166	Wor.	0.148
Weather	Loc.	0.336	Dis.	0.199	Liv.	0.173
Working	Exp.	0.279	Loc.	0.218	Con.	0.191

$$\mathrm{F-measure}(a) = \frac{2 * \mathrm{Precision}(a) * \mathrm{Recall}(a)}{\mathrm{Precision}(a) + \mathrm{Recall}(a)}, \tag{4.10}$$

where D is the number of tweets for evaluation.

We judged the experimental evaluations by 10-fold cross validation. We split the datasets into 10 subsets, only one of which is circularly selected as a test dataset. We built associations between topics and the aspects using the remaining nine subsets.

4.2.3 Baseline methods

We prepared such typical multi-label classification methods as L-LDA [58], LIBSVM [12], and NBML [76] for evaluating HEF's effectiveness with the entropy optimizations. We extracted nouns, verbs, and adjectives using a Japanese morphological analyzer called MeCab [37] and entered the sets of words and label(s) to every method in common.

LIBSVM requires some parameters. We chose a linear kernel and set parameter C to 1.0, indicated by a grid search in the LIBSVM tools[24]. The features for all the methods are nouns, verbs, and adjectives, which were obtained by morphological analysis.

L-LDA has to set the hyperparameters of both α and β, like in LDA. We experimentally set α to 0.1 and β to 0.1, and the iterative calculation count in L-LDA was 100.

4.2.4 Experimental results

Number of topics

We evaluated JS_{sum} to tune the number of topics. The list of JS_{sum} that varies the number of topics is shown in **Fig. 4.1**. The maximum value appears in 500 topics. We concurrently evaluated the precision, the recall, and the F-measure in each topic. The maximum precision and recall were achieved in 200 and 1,000 topics, and the maximum F-measure was achieved in 500 topics. Therefore, we used 500 as the optimal number of topics for HEF. The decision method of the optimal number of topics by the JS_{sum} value is generally effective for HEF because stable evaluation values were achieved in about 500 topics.

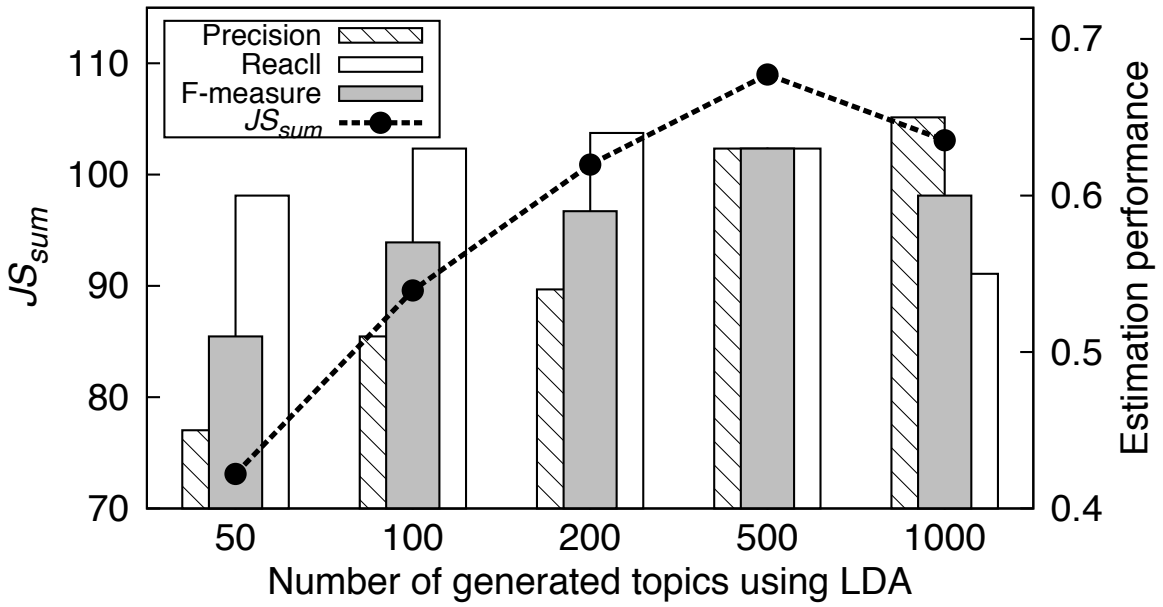

Figure 4.1: JS_{sum}, Precision, Recall, and F-measure values of each number of topics

Feedback coefficients

The feedback coefficients of both α and β vary, as shown in **Fig. 4.2**. The convergence condition was set within a difference of 0.00001 compared to previous iteration values. The starting value was set to 1.0. Both converged at eight iterations. From this result, α and β became 0.85626 and 0.22655.

Threshold parameters

To confirm the automatically decided threshold $r(a)$ of Eq. (4.5), we evaluated the precision, recall, and F-measure values of our proposed method by three simple thresholds: $r = 0.0$, $r = 0.5$, and $r = 1.0$. The feedback coefficients have identical values in all the thresholds.

The precision, recall, and F-measure values by each threshold are shown in **Table 4.3**. The maximum precision and recall values were achieved by $r = 1.0$ and $r = 0.0$. However, the minimum recall and precision values also are demonstrated by these thresholds. $r(a)$, which automatically decided the threshold, achieved the highest F-measure value in all the thresholds. This result shows that our proposed threshold

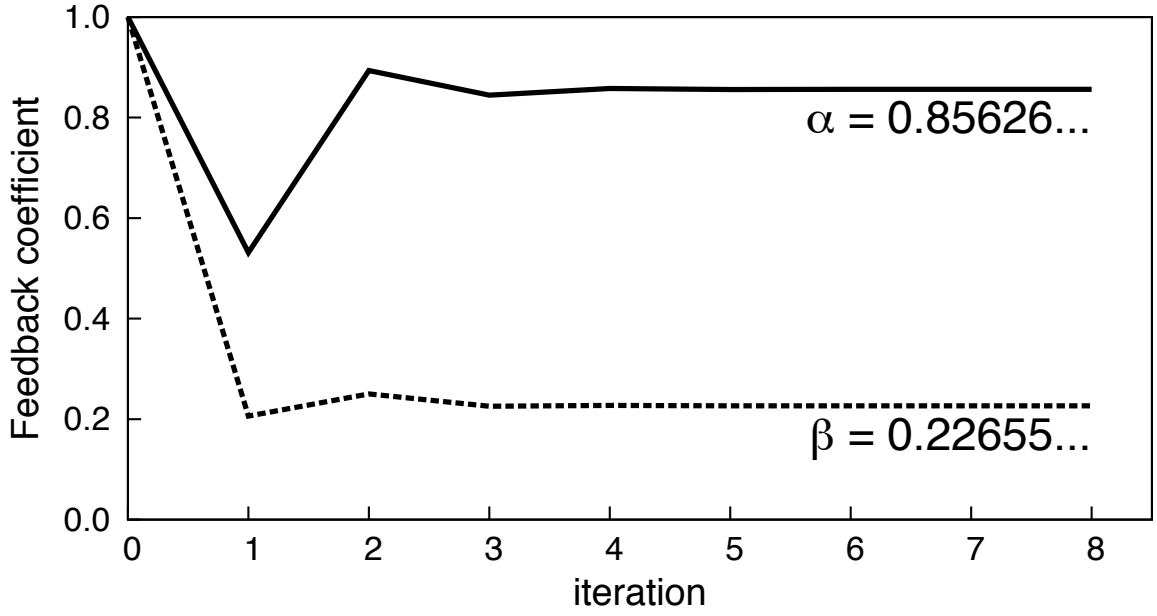

Figure 4.2: Converging state of feedback coefficients

Table 4.3: Compared to estimation performance of HEF using each threshold value

Threshold	Precision	Recall	F-measure
$r = 0.0$	0.57	**0.66**	0.59
$r = 0.5$	0.61	0.57	0.58
$r = 1.0$	**0.68**	0.54	0.58
$r(a)$	0.63	0.63	**0.63**

is effective for HEF.

Connections from topics to each aspect

To analyze the association between topics and aspects, we evaluated the number of connections from the topics to each aspect. The number of topics connecting each aspect varying to parameter d is shown in **Fig. 4.3**. In all aspects, the number of topics increased based on d. The Appearance aspect is most closely connected to one topic, $d \leq 11$. The Hobby aspect connects to much topics with fewer value of d, and it completely connects to all the topics at d=6. When d exceeds 18, the associations

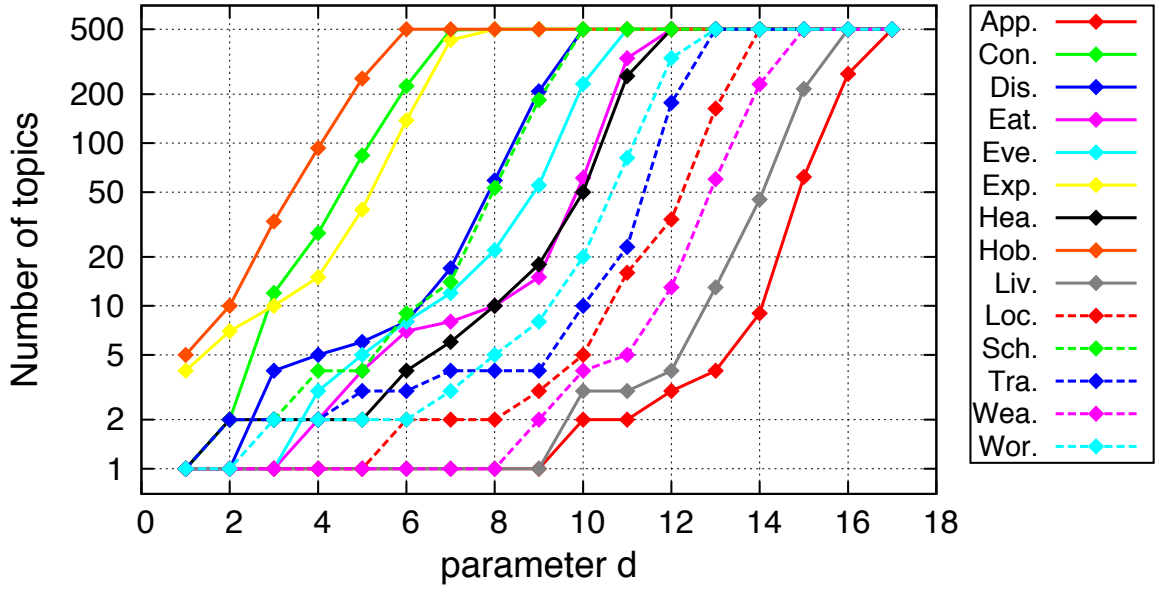

Figure 4.3: Connectivity among topics and aspects

between topics and aspects become a complete bipartite graph.

Transitions of precision, recall, and F-measure values

The precision, recall, and F-measure values are shown in from **Fig. 4.4** to **Fig. 4.18** for all aspects. The horizontal axis is parameter d, which decides the association between topics and aspects.

In the Disaster aspect, recall slowly increased based on d. In contrast, precision decreased based on d. The maximum F-measure was achieved at $d=5$.

In the School aspect, precision rapidly increased until $d \leq 6$ and then quickly decreased until $d \leq 9$. Recall decreased until $d \leq 4$ and then increased until $d \leq 9$. The maximum F-measure was achieved at $d=6$.

In the Traffic aspect, the precision, recall, and F-measure values increased until $4 \leq d \leq 6$. There are three topics at $d=6$. Precision increased until $d \leq 10$ and then decreased until $d \leq 14$. The maximum F-measure was achieved at $d=10$.

The evaluation values change even after they are connected to all the topics in these aspects. The associations of other aspects change based on an increased d until

$d \leq 18$, and the aspect scores also change. We show the optimal d of each aspect in the next section.

Estimation performance of each method

The precision, recall, and F-measure values of each method are shown in **Table 4.4, 4.5, and 4.6**. All of the methods were evaluated using 10-fold cross validations. In each evaluation, 1,350 tweets were used for model training, and the remaining 150 tweets were used to evaluate the precision, recall, and F-measure values. We also calculated their macro averages. The highest value in each row is shown in bold. The HEF columns show our method, where associations were built using entropy feedback. The HEF0 columns show the 0 iteration cases of entropy feedback, and both α and β are 1.0. Optimal values of d and number of topics when achieved the highest F-measure are shown in the d and $|T_a|$ columns in **Table 4.1**. Optimal d of Appearance is 17 when precision and recall are 0.74 and 0.53. In the Disaster and Traffic aspects, HEF's precisions greatly increased without decreasing recall more than HEF0's. Disaster's F-measure by HEF surpassed 0.25 points ($= 0.54 - 0.29$) compared with HEF0. HEF's average F-measure showed the highest value in all the methods.

The number of labels, each of which was estimated as an aspect of the tweets by all methods and the examinees, is shown in **Table 4.7**. In the examinees and SVM, there are three labeling modes. The maximum and minimum numbers of labeling modes in every method are found in the HEF and NBML values.

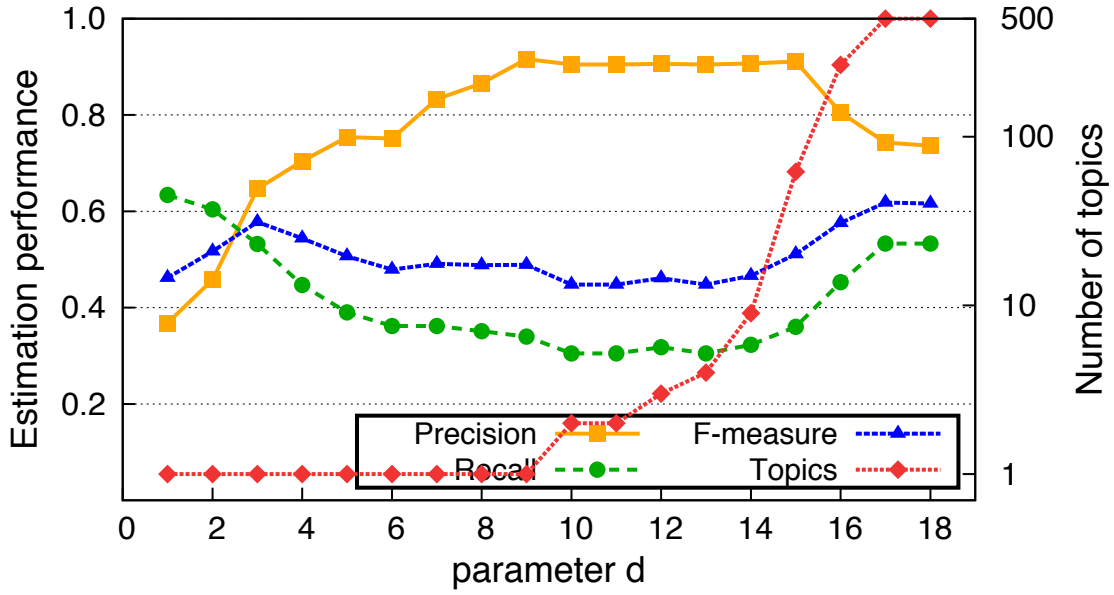

Figure 4.4: Precision, Recall, and F-measure of Appearance

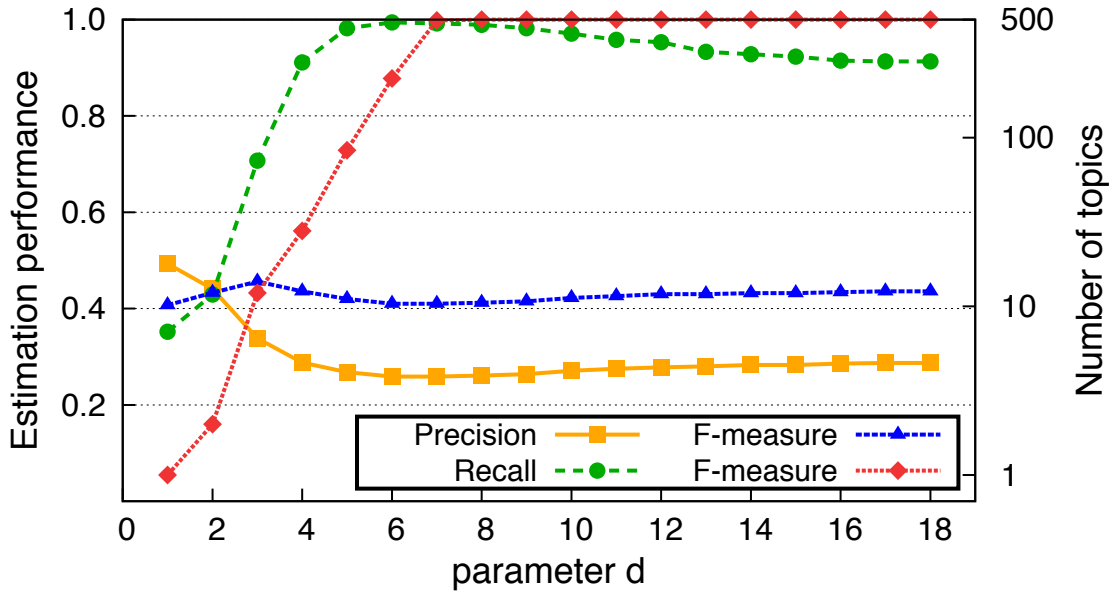

Figure 4.5: Precision, Recall, and F-measure of Contact

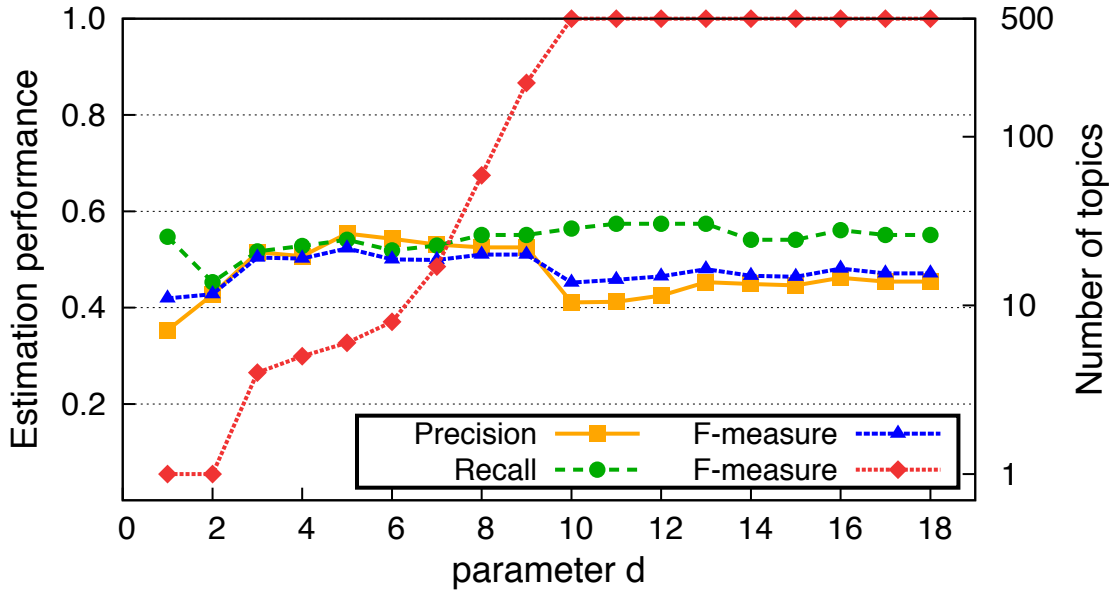

Figure 4.6: Precision, Recall, and F-measure of Disaster

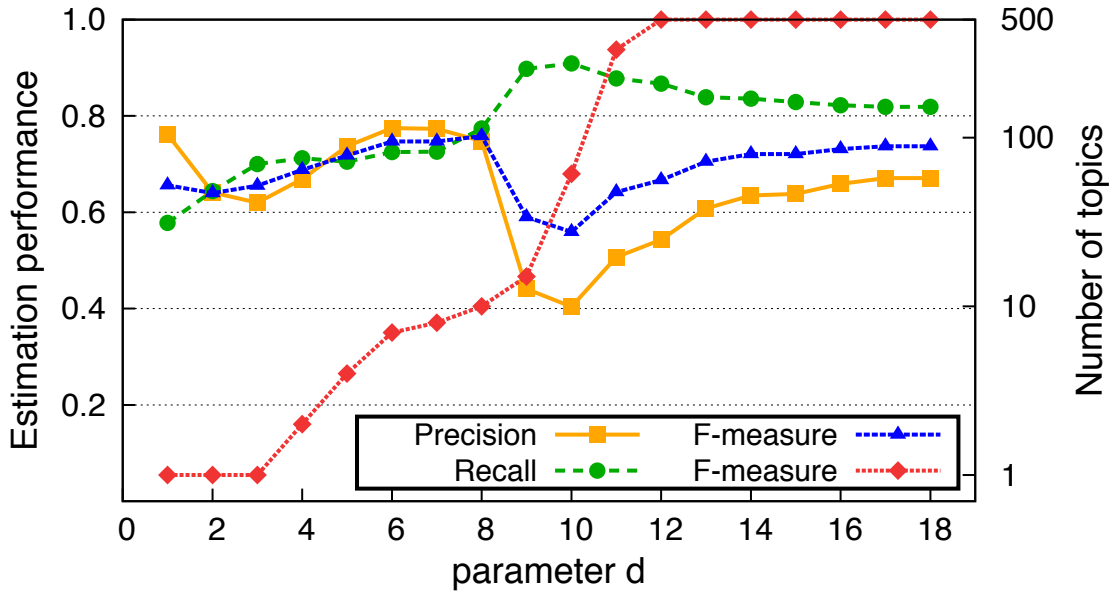

Figure 4.7: Precision, Recall, and F-measure of Eating

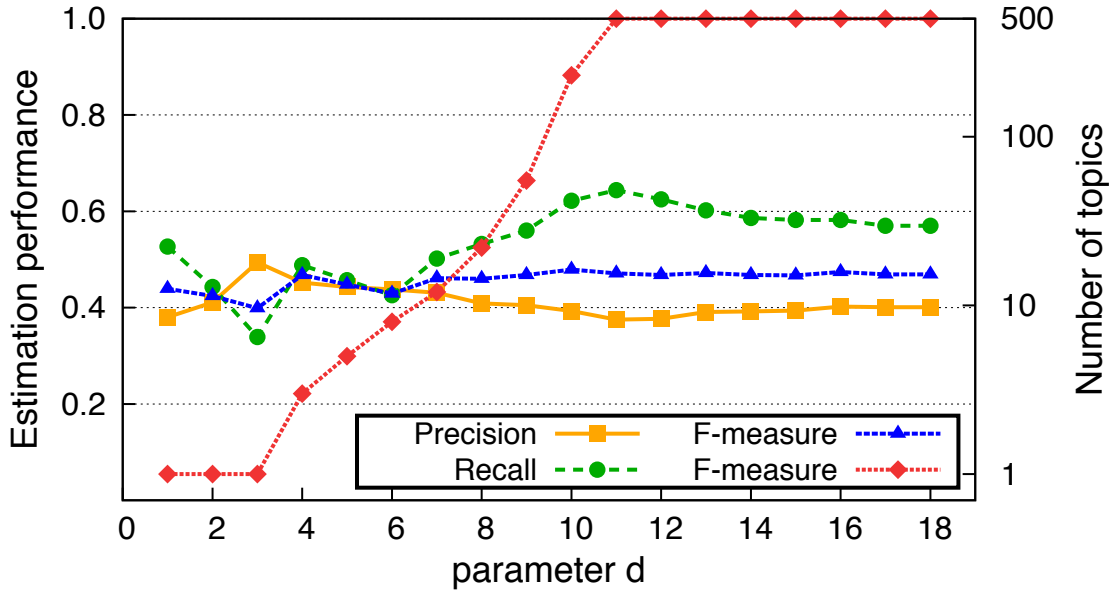

Figure 4.8: Precision, Recall, and F-measure of Event

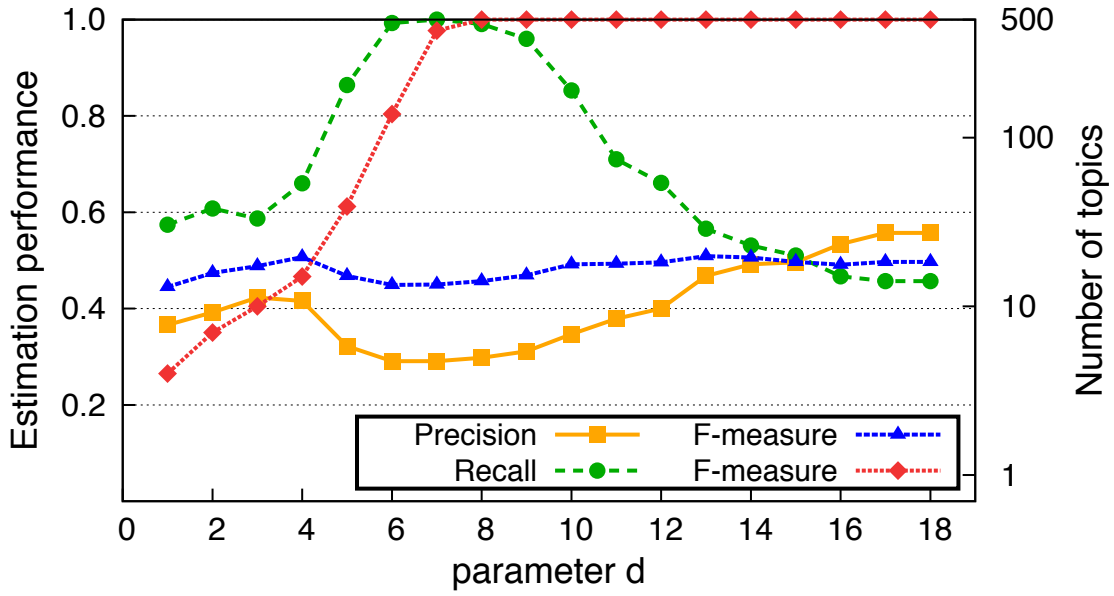

Figure 4.9: Precision, Recall, and F-measure of Expense

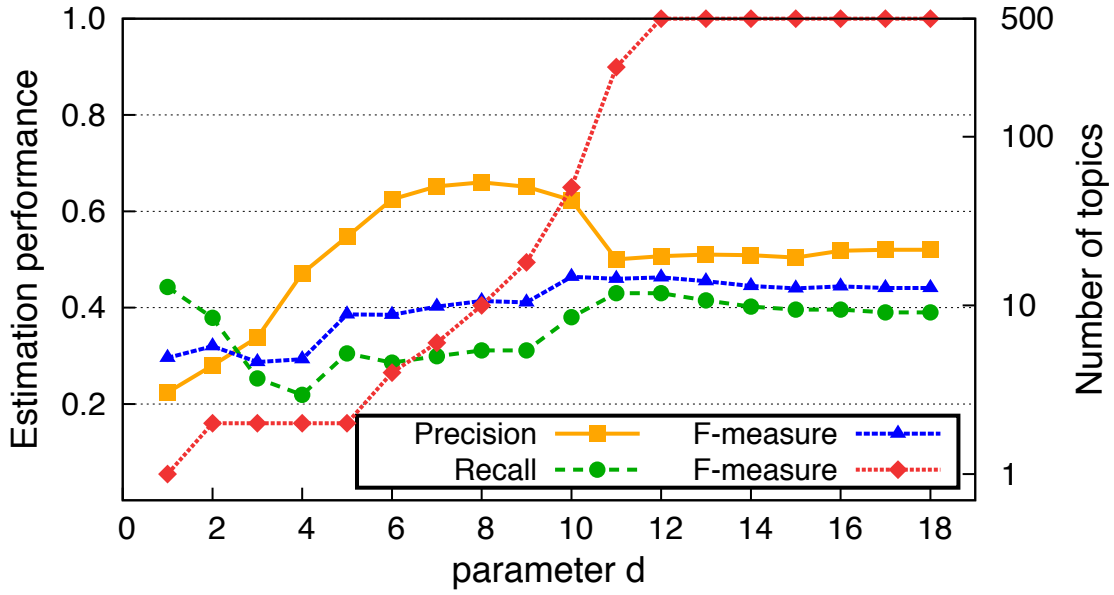

Figure 4.10: Precision, Recall, and F-measure of Health

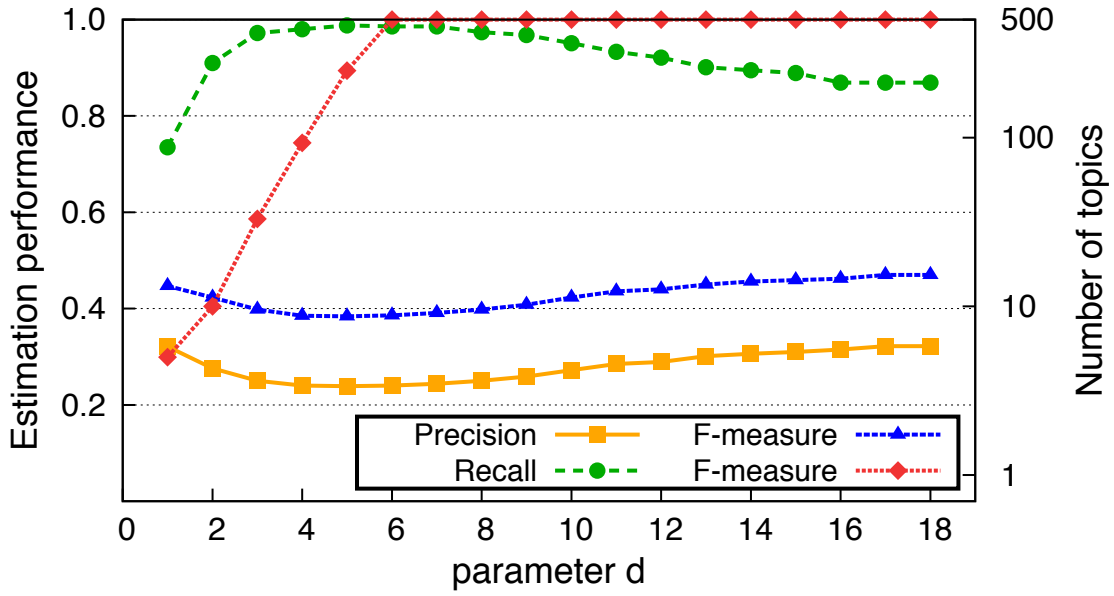

Figure 4.11: Precision, Recall, and F-measure of Hobby

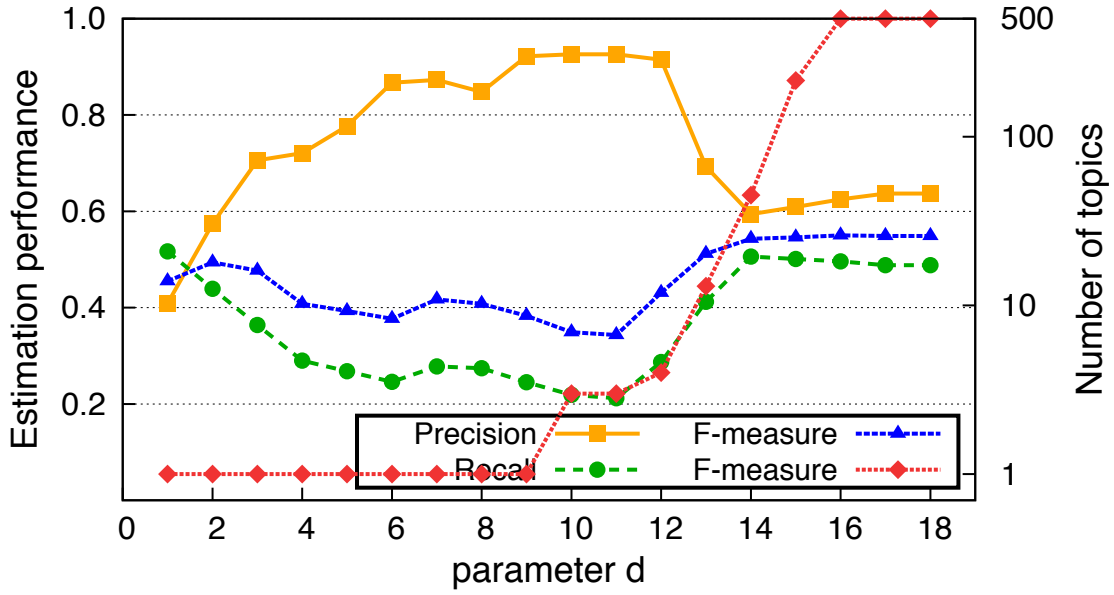

Figure 4.12: Precision, Recall, and F-measure of Living

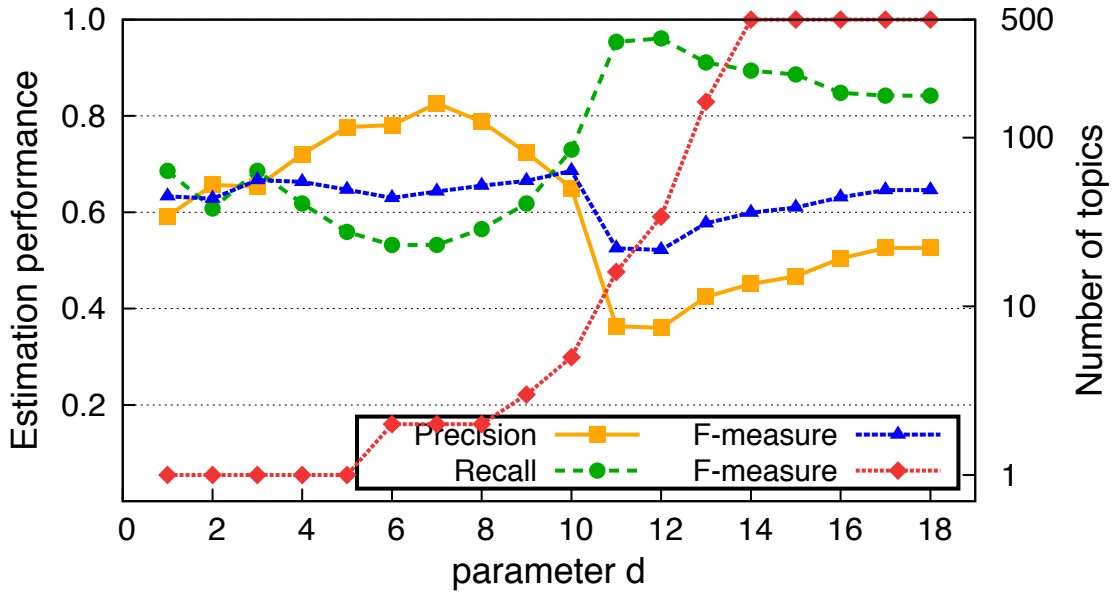

Figure 4.13: Precision, Recall, and F-measure of Locality

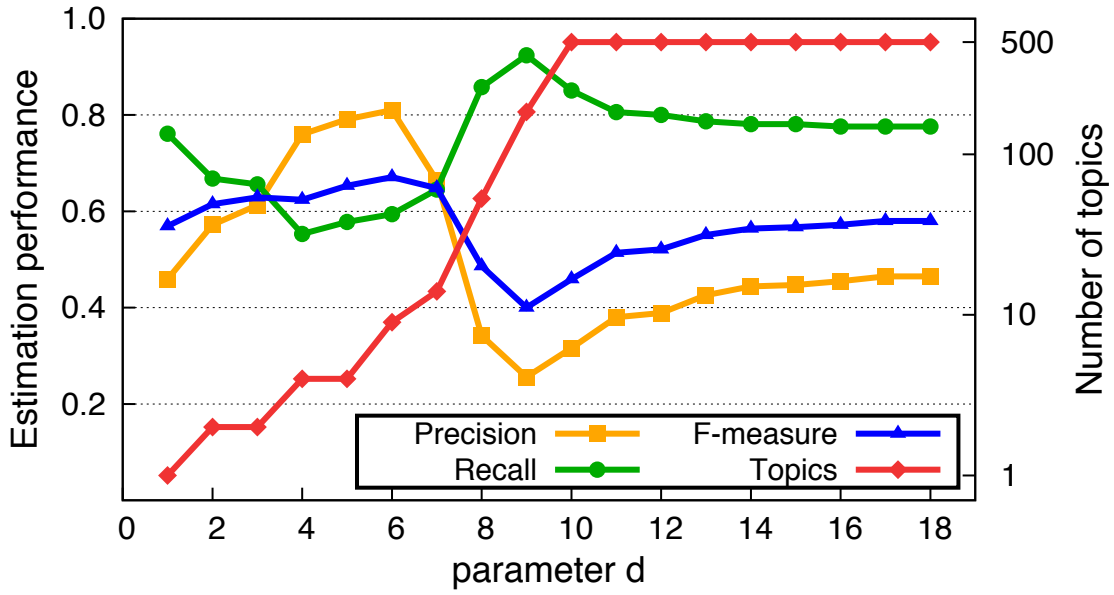

Figure 4.14: Precision, Recall, and F-measure of School

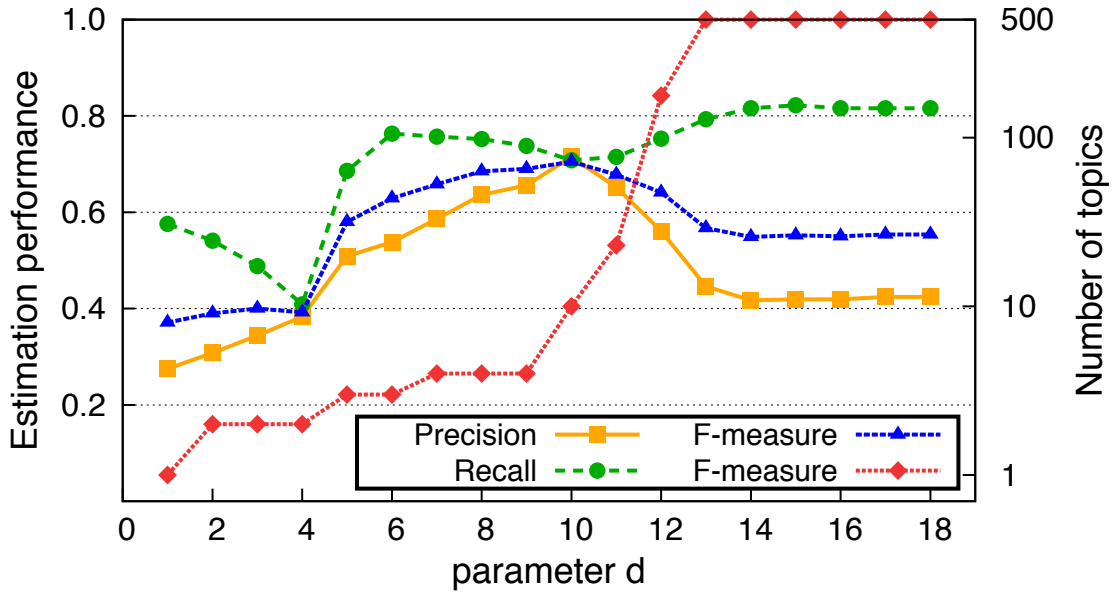

Figure 4.15: Precision, Recall, and F-measure of Traffic

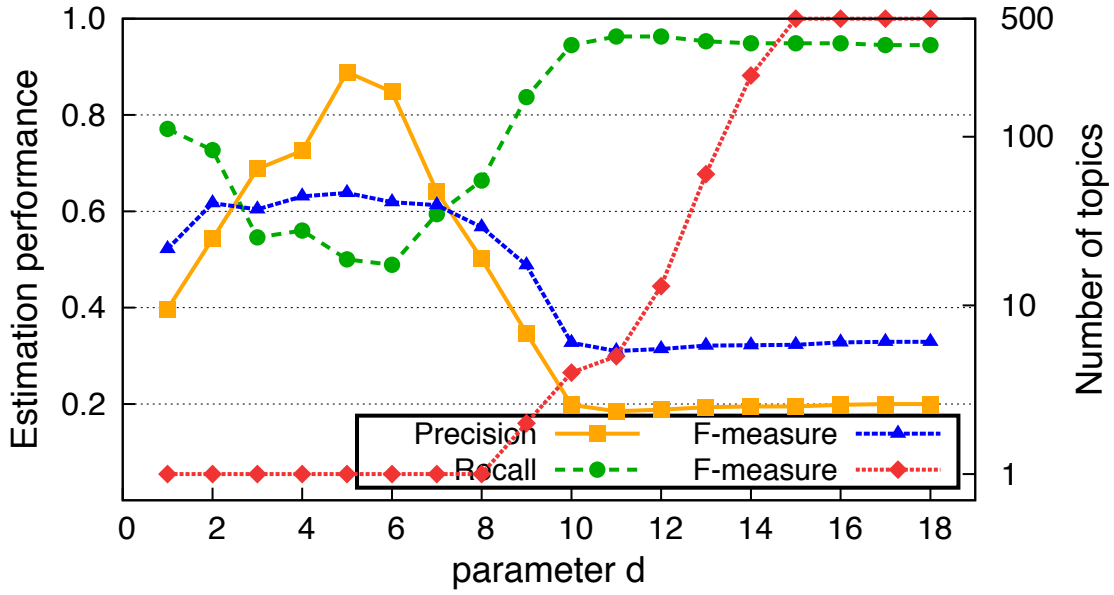

Figure 4.16: Precision, Recall, and F-measure of Weather

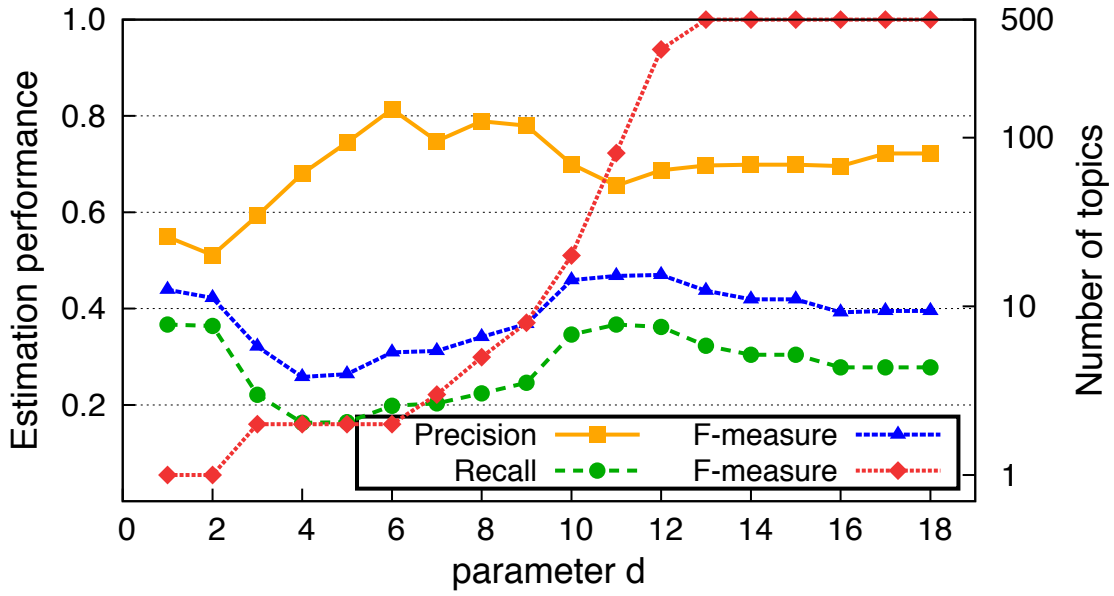

Figure 4.17: Precision, Recall, and F-measure of Working

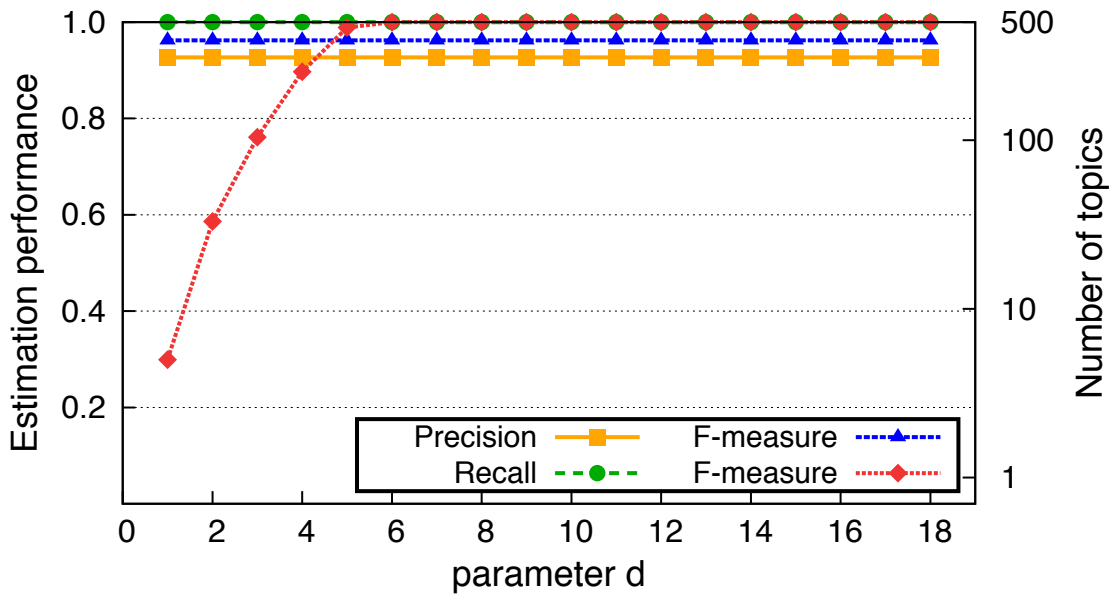

Figure 4.18: Precision, Recall, and F-measure of Other

Table 4.4: Precision of each method

	Precision				
Aspect	HEF0	HEF	L-LDA	SVM	NBML
Appearance	0.61	0.74	0.43	0.64	**0.82**
Contact	0.40	0.34	0.43	0.41	**0.53**
Disaster	0.21	0.55	0.67	0.44	**0.76**
Eating	0.66	**0.75**	0.41	0.51	0.73
Event	0.41	0.39	0.47	**0.56**	**0.56**
Expense	0.39	0.47	**0.64**	0.43	0.52
Health	0.31	0.62	0.43	0.48	**0.76**
Hobby	0.32	0.32	0.44	0.43	**0.57**
Living	0.34	0.63	0.38	0.64	**0.71**
Locality	**0.65**	**0.65**	0.62	0.62	0.62
School	0.57	0.81	0.37	**0.88**	0.81
Traffic	0.54	0.72	0.33	0.71	**0.82**
Weather	0.28	**0.89**	0.25	0.47	0.81
Working	0.38	**0.69**	0.64	0.52	0.56
Other	0.93	0.93	**0.94**	0.93	0.93
Average	0.47	0.63	0.50	0.58	**0.70**

Table 4.5: Recall of each method

	Recall				
Aspect	HEF0	HEF	L-LDA	SVM	NBML
Appearance	0.54	0.53	**0.69**	0.28	0.37
Contact	0.55	**0.71**	0.68	0.35	0.54
Disaster	0.49	**0.54**	0.49	0.44	0.21
Eating	0.74	**0.77**	**0.77**	0.64	0.51
Event	0.55	**0.62**	0.51	0.20	0.45
Expense	**0.65**	0.57	0.40	0.45	0.46
Health	**0.56**	0.38	0.55	0.28	0.38
Hobby	0.84	**0.87**	0.62	0.54	0.44
Living	**0.74**	0.50	0.62	0.34	0.41
Locality	0.66	**0.73**	**0.73**	0.54	0.65
School	0.59	0.59	**0.81**	0.36	0.52
Traffic	0.68	0.71	**0.82**	0.44	0.50
Weather	0.81	0.50	**0.84**	0.63	0.58
Working	**0.64**	0.36	0.50	0.19	0.35
Other	**0.99**	**0.99**	0.51	**0.99**	0.93
Average	**0.67**	0.63	0.63	0.44	0.49

Table 4.6: F-measure of each method

	F-measure				
Aspect	HEF0	HEF	L-LDA	SVM	NBML
Appearance	0.57	**0.62**	0.52	0.38	0.51
Contact	0.46	0.46	**0.53**	0.37	**0.53**
Disaster	0.29	**0.54**	**0.54**	0.44	0.33
Eating	0.70	**0.76**	0.53	0.57	0.60
Event	0.47	0.48	0.48	0.29	**0.49**
Expense	0.49	**0.51**	0.49	0.43	0.49
Health	0.40	0.46	0.48	0.35	**0.50**
Hobby	0.46	0.47	**0.51**	0.48	0.49
Living	0.46	**0.55**	0.46	0.44	0.51
Locality	0.65	**0.69**	0.67	0.57	0.63
School	0.58	**0.67**	0.51	0.49	0.63
Traffic	0.60	**0.71**	0.47	0.54	0.62
Weather	0.41	0.64	0.38	0.53	**0.67**
Working	0.47	0.47	**0.55**	0.28	0.43
Other	**0.96**	**0.96**	0.66	**0.96**	0.93
Average	0.55	**0.63**	0.52	0.47	0.56

Topics associated with each aspect

Next we examined the topics connected to each aspect. The associations built by HEF0 are shown in **Table 4.8**. This table shows the topic ids of top four that are strongly connected to each aspect. Three or more times appearing topics in every aspect are marked in bold. For example, the Appearance aspect is associated to topic 119 with highest relevance $\hat{Ra}(a,t)$. Topic 125 associates to the Disaster, Event, Locality, and Traffic with the highest (1st rank) relevances and Weather with the second highest (2nd rank) relevance. Moreover, topics 125, 299, and 469 appear

Table 4.7: Number of labelings by each method

Labels	HEF	L-LDA	SVM	NBML	Examinees
1	0	101	0	165	1
2	0	154	137	531	111
3	22	259	1,250	442	820
4	389	369	80	243	442
5	574	307	33	90	115
6	382	182	0	23	11
7	118	80	0	6	0
8	15	37	0	0	0
9	0	9	0	0	0
10	0	2	0	0	0
Average	5.15	4.16	3.00	2.75	3.39

together in the Disaster, Event, and Locality aspects. Topic 60 appears in the aspects of Disaster, Locality, and Traffic.

Similarly, the associations built by HEF are shown in **Table 4.9**. Note that these topic ids are same as HEF0's topic ids. The associations built by HEF are quite different from those by HEF0. For example, topic 125 appears only in the Locality aspect with 4th rank unlike in HEF0's associations. The aspects of Disaster and Event are associated topic 178 and 345 with 1st rank. Topic 60 connects to both aspects of Locality and Traffic with 1st rank.

Table 4.8: Relevance of high $\hat{Ra}$ topics built by HEF0

	$\hat{Ra}$ Rank 1			$\hat{Ra}$ Rank 2			$\hat{Ra}$ Rank 3			$\hat{Ra}$ Rank 4		
Aspect	topic	$\hat{Ra}$	$\hat{Rt}$	topic	$\hat{Ra}$	$\hat{Rt}$	topic	$\hat{Ra}$	$\hat{Rt}$	topic	$\hat{Ra}$	$\hat{Rt}$
Appearance	#119	0.099	0.981	#368	0.048	0.204	#458	0.044	0.199	#164	0.043	0.206
Contact	# 49	0.039	0.560	# 9	0.033	0.466	#157	0.032	0.236	#490	0.023	0.155
Disaster	**#125**	0.196	0.189	**#299**	0.076	0.179	**#469**	0.064	0.191	**# 60**	0.052	0.155
Eating	#207	0.069	0.771	#197	0.068	0.758	#484	0.039	0.790	#352	0.038	0.766
Event	**#125**	0.135	0.179	#345	0.053	0.589	**#299**	0.053	0.171	**#469**	0.044	0.177
Expense	#437	0.058	0.458	#454	0.054	0.443	# 11	0.027	0.164	#223	0.022	0.166
Health	#237	0.075	0.205	#359	0.064	0.200	#479	0.055	0.196	#393	0.055	0.847
Hobby	#221	0.048	0.629	#332	0.040	0.630	#311	0.033	0.621	#497	0.031	0.617
Living	#290	0.062	0.948	#275	0.046	0.309	#301	0.039	0.752	# 11	0.031	0.228
Locality	**#125**	0.122	0.267	**#299**	0.050	0.269	**# 60**	0.041	0.280	**#469**	0.039	0.262
School	# 3	0.066	0.857	#490	0.026	0.126	#443	0.025	0.142	#275	0.025	0.195
Traffic	**#125**	0.119	0.170	**# 60**	0.054	0.239	**#299**	0.053	0.186	#201	0.045	0.871
Weather	#451	0.044	0.390	**#125**	0.035	0.138	#490	0.033	0.313	#230	0.029	0.582
Working	#334	0.068	0.611	#253	0.065	0.471	# 21	0.060	0.194	#463	0.058	0.196
Other	#237	0.020	0.162	#359	0.018	0.167	#479	0.016	0.169	#490	0.015	0.131

Table 4.9: Relevance of high $\hat{Ra}$ topics built by HEF

	$\hat{Ra}$ Rank 1			$\hat{Ra}$ Rank 2			$\hat{Ra}$ Rank 3			$\hat{Ra}$ Rank 4		
Aspect	topic	$\hat{Ra}$	$\hat{Rt}$	topic	$\hat{Ra}$	$\hat{Rt}$	topic	$\hat{Ra}$	$\hat{Rt}$	topic	$\hat{Ra}$	$\hat{Rt}$
Appearance	#119	0.029	0.668	#474	0.010	0.358	#240	0.007	0.314	#454	0.006	0.147
Contact	# 49	0.009	0.312	#429	0.008	0.192	#157	0.007	0.177	#466	0.006	0.140
Disaster	#178	0.021	0.271	#380	0.017	0.382	**#469**	0.014	0.215	#277	0.012	0.405
Eating	#341	0.022	0.776	#484	0.018	0.537	#352	0.017	0.577	#207	0.016	0.529
Event	#345	0.026	0.320	#314	0.018	0.225	#190	0.017	0.300	#307	0.015	0.217
Expense	#437	0.010	0.281	# 35	0.010	0.210	#454	0.009	0.228	#419	0.009	0.194
Health	#393	0.022	0.468	# 22	0.021	0.591	#348	0.013	0.259	#193	0.013	0.417
Hobby	# 75	0.007	0.206	#412	0.007	0.430	#273	0.007	0.170	#430	0.007	0.362
Living	#290	0.022	0.520	#133	0.012	0.349	#230	0.011	0.270	#301	0.007	0.322
Locality	**# 60**	0.022	0.282	#314	0.017	0.264	**#299**	0.011	0.199	**#125**	0.010	0.233
School	# 3	0.015	0.464	#111	0.014	0.582	#118	0.011	0.519	#418	0.010	0.329
Traffic	**# 60**	0.031	0.342	#201	0.022	0.588	#149	0.019	0.468	# 42	0.014	0.403
Weather	# 23	0.020	0.441	#451	0.013	0.292	#490	0.009	0.236	#178	0.009	0.212
Working	#321	0.017	0.466	#436	0.014	0.258	#253	0.010	0.220	#334	0.009	0.215
Other	#281	0.005	0.125	#330	0.005	0.147	#304	0.005	0.150	# 21	0.005	0.129

Table 4.10: High occurrence probability terms in each topic associated to many aspecs

Topic	Characteristic words
Topic 125	Kyoto, newspapers, city centers, aquariums, towers, Yamashina, Sakyo, living, Fushimi
Topic 299	Kyoto, sightseeing, hotels, taxis, roaming, travel, school trips, lodging
Topic 469	Kyoto, citizens, institutions, environments, nursing, welfare, newspapers, medical
Topic 490	today, work hard, tomorrow, energy, hot, part-timer, work, sunny

Estimation precision using a small bit of labeled data

In all the methods, we evaluated the estimation performance using less training data. We split the datasets into 10 subsets, and only one subset is circularly selected as a test dataset. From the remaining nine subsets, we randomly extracted 1 set (150 tweets), 3 sets (450), 5 sets (750), and 7 sets (1050) as the training data. We calculated the average evaluation value by repeating ten times changing the test data. Each evaluation value is shown in **Fig. 4.19**, **4.20**, and **4.21**. We chose optimal d as the HEF parameters. The optimal number of the topics in all the training data was 500, depending on JS_{sum}.

HEF's precision is lower than NBML's with training dataset 9 (1350). However, based on the decreasing training data, the precision difference of both methods was small. The precision of our method did not fall even when the amount of training data decreased. In recall, the precision of L-LDA and SVM rapidly fell with less training data; however, HEF and NBML showed almost no drop. In the F-measures, HEF achieved the high score until training dataset 3, and it is usually the maximum F-measure in all the methods. The F-measure of SVM rapidly dropped with less training data.

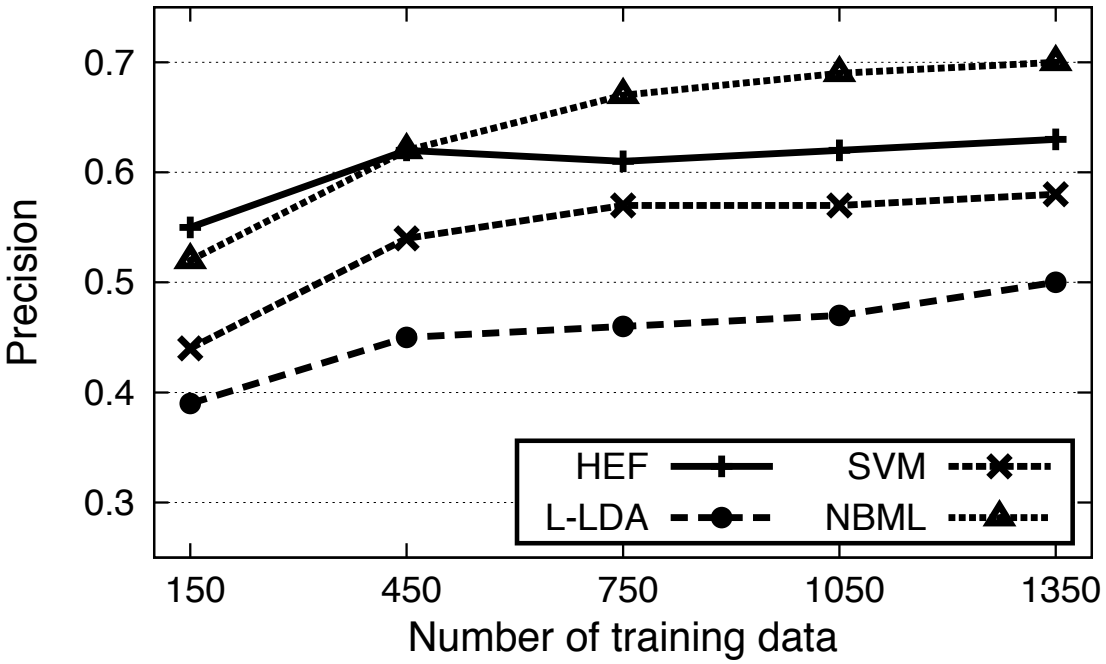

Figure 4.19: Precision evaluated by varying amount of training data

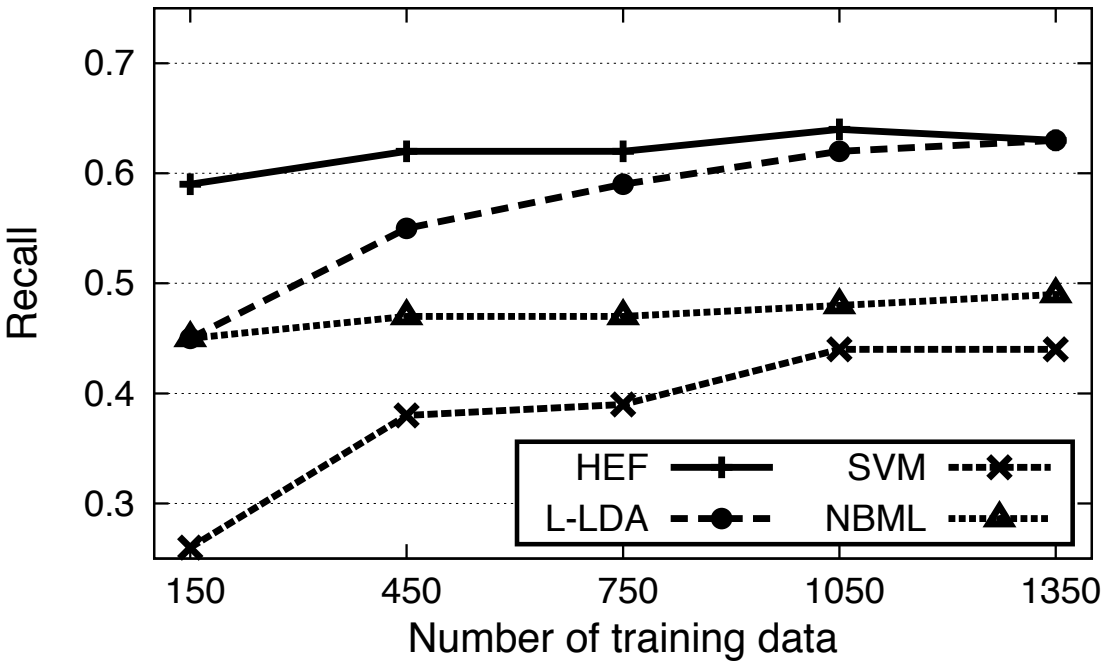

Figure 4.20: Recall evaluated by varying amount of training data

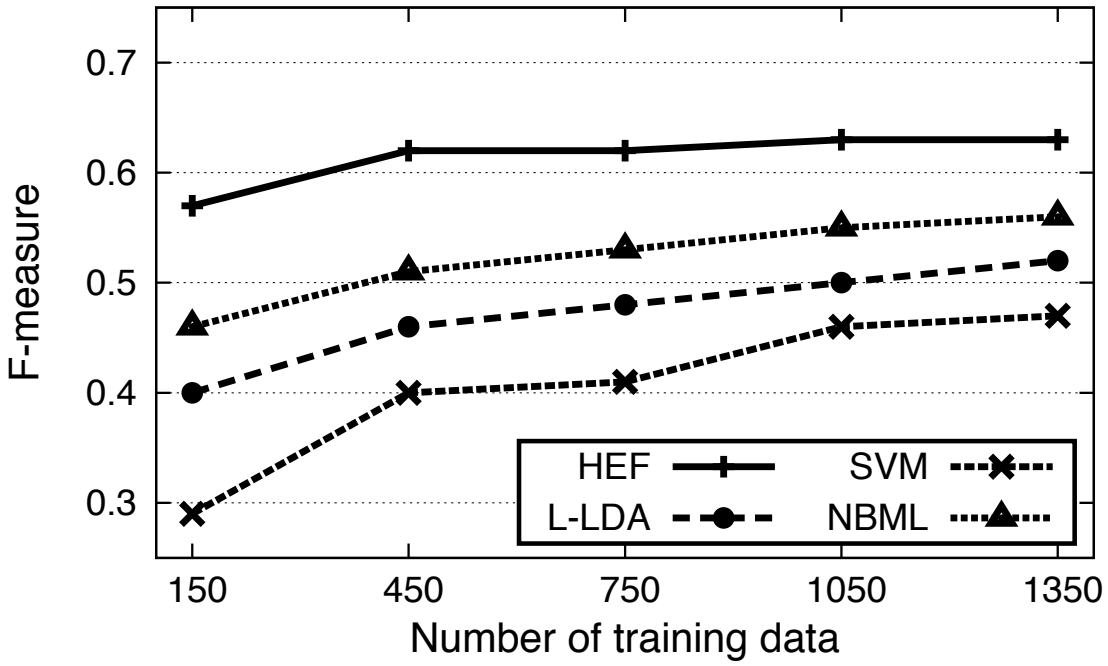

Figure 4.21: F-measure evaluated by varying amount of training data

4.3 Discussions

4.3.1 Effectiveness of feedback entropy

According to **Table 4.6**, HEF's F-measure increased more than HEF0's F-measure in many aspects, especially for Disaster, Traffic, and Weather F-measure values. From **Table 4.8**, Disaster is strongly associated to Topics 125, 299, 460, and 60. Part or all of them were also associated with Event, Locality, Traffic, and Weather aspects. The characteristic words in HEF0 are shown in **Table 4.10**. Topic 125 has "aquarium" and "tower" that denote the names of structures, and "Yamashina" and "Sakyo" denote place names. Topic 125 is related to the names of geographic elements around Kyoto. Topic 299 is related to tourism/tourists in Kyoto because it includes "sightseeing" and "travel." Topic 469 is related to living in Kyoto because "welfare" and "nursing" are found in it. In these topics, "Kyoto" exists at the top priority of the characteristic words. Therefore, these topics describe the Kyoto district and are

connected to the Locality aspect; however, they are also connected to other aspects, such as Disaster, Event, and Traffic. To explain these diverse connections, Disaster or Event tweets frequently include such place words. For example, earthquake tweets usually describe not only the earthquake itself but also its center, which is obviously a geographic name. In topics about districts, geographic names have very high occurrence probability, as shown in **Table 4.10**. For these reasons, similar sets of topics are connected to the Disaster, Event, and Traffic aspects.

On the other hand, for the associations built by HEF, almost none of these topics appeared in all of the aspects (**Table 4.9**). The most strongly associated topics with Disaster, Locality, Traffic, and Weather are Topics 178, 60, and 23, respectively. Their characteristic words are shown in **Table 4.11**. Topic 178 has "typhoon" and "storm," which denote natural disasters, and the tweet was associated with a Weather aspect in $\hat{Ra}$ rank 4. Topic 60 has such place names as "Kyoto" and "Kawaramachi." It also has "subway" and "city bus," which are usually used for the Traffic aspect. For these reasons, the Locality and Traffic aspects share Topic 60. Topic 23 has "sunny" and "forecast," which are usually used for the Weather aspect. These relationships among characteristic words and real life aspects are also shown in **Table 1.1**. From these results, higher F-measures in the aspects of Disaster, Locality, Traffic, and Weather are achieved by strongly connected topics: Topics 178, 60, and 23.

4.3.2 Estimation performance of each method

From **Table 4.4**, **4.5**, and **4.6**, HEF's average precision (0.63) is lower than NBML's. But HEF's average recall (0.63) and its average F-measure (0.63) are higher than NBML's. In **Table 4.7**, the number of labelings by NBML is the lowest in all the methods and fewer than the labels of the examinees. NBML's precision rose but its recall fell. When we compare the average recalls of HEF and L-LDA, we see that HEF's recall is the same as L-LDA's. From **Table 4.7**, HEF estimates more labels than L-LDA. However, its precision and its F-measure are higher than L-LDA's. Since

HEF estimated more correct aspects than L-LDA, our method accurately calculated the aspect scores of tweets. To enhance HEF's precision, we introduce into Eq. (4.5) an adjustment parameter, which is a threshold that controls the number of labels. We increase the threshold values with all the aspects by a parameter. HEF estimates the aspects with higher scores, and the number of labels by HEF approaches to the number of human labelings. The optimal threshold value, which approaches the number of human labelings, is obtained by calculating the parameter regarding it.

From **Fig. 4.19, 4.20**, and **4.21**, our method shows the results where the descent of the precision is small. The recalls of HEF and L-LDA have almost no difference with training dataset 9. With less training data, L-LDA's recall rapidly dropped. However, HEF showed almost no drop. In every method except HEF, the recall values rapidly fell because the terms decreased based on less training data. On the other hand, in our method, topics are associated to aspects. Therefore, the terms don't decrease even if the number of training data decreased. For these reasons, the HEF's recall almost didn't fall based on less training data. Hence, the HEF's F-measure is higher than all the other methods.

A sample tweet is shown in **Table 4.12**. The examinee aspect column shows the aspects labeled by the examinees, based on Eq. (4.7). The columns of the HEF, HEF0, and NBML aspects estimated the aspects by each method. **Table 4.12** is a completely matched example whose aspects were labeled by the examinees and estimated by HEF. It shows the effectivity and the characteristics of HEF estimation and mentions a restaurant's opening in "Takaragaike". The examinee aspects are Eating, Expense, and Locality, all of which coincide with the tweet's topics. NBML estimated Eating and Expense aspects but failed to estimate the Locality aspect because it was not trained by the likelihood between "Takaragaike" and Locality by the training data. HEF0 estimates many aspects: Eating, Expense, Locality, Disaster, Event, and Traffic. Obviously, this tweet does not mention Traffic, Disaster, or Event. HEF0 excessively estimated aspects because it built associations between

many aspects and local topics, such as Topic 125. On the other hand, the aspects estimated by HEF match the examinees' aspects. HEF estimated Locality because it built associations between the Locality aspect and a topic including "Takaragaike". Since other aspects were not associated to the local topics, HEF accurately estimated the aspects.

Table 4.13 is an example where our prototype system estimated other aspects, in addition to the aspects of the examinees. This tweet mentions conducting a workshop. The examinee aspects are School and Event, and they coincide with the tweet's topic. The tweet includes the term "Katsura-River" which is the name of a river in Kyoto. Because this tweet is about an event being held in the "Katsura-River" neigborhood, it includes the Locality aspect which our prototype system can estimate. This sample is a good example that our proposed method effectively identified.

Table 4.11: High occurrence probability terms in each topic strongly associated to each aspect

Topic 119 (Appearance)	wear, Yukata, t-shirt, one-piece, suit, costume, uniform, Kimono
Topic 49 (Contact)	fun, everyone, drink, partying, wild, yesterday, meet, party
Topic 178 (Disaster)	typhoons, Kyoto, alarm, heavy rain, storms, influence, precautions, floods
Topic 341 (Eating)	curry, vegetable, tomato, delicious, cook soup, salad, sauce, eat
Topic 345 (Event)	participation, event, hold, plan, detail, offer, decision, member
Topic 437 (Expense)	buy, cheap, price, supermarket convenience store, half price, used, sell
Topic 393 (Health)	hospital, expert, outpatient clinic, health treatment, diagnosis, examination, causation
Topic 75 (Hobby)	watch, think, interesting, uninteresting, miracle, animation, theater, boy
topic 290 (Living)	room, cleaning, open, dirty clear, toilet, close, door
Topic 60 (Locality and Traffic)	Kyoto, traffic, Kawaramachi, Shijyo, Torimaru, subways, guides, city buses
Topic 3 (School)	study, exam, finish, concentration period, practice exam, subject, score
Topic 23 (Weather)	weather, sunny, forecast, Kyoto, rainy season, clouds, temperature
Topic 321 (Working)	company, corporation, finding employment, employee company president, management, career-change

Table 4.12: Complete estimated aspects for tweet by HEF

Answers	Locality, Expense, Eating
HEF	Locality, Expense, Eating
HEF0	Locality, Expense, Eating, Event, Disaster, Traffic
NBML	Expense, Eating
Tweet	Any plans for the weekend? How about some curry? We're opening a new curry restaurant in front of the Takaragaike baseball stadium on the 24th.

Table 4.13: Estimated extra aspects for tweet

Answers	School, Event
HEF	School, Event, Locality
NBML	School, Event
Tweet	Attention! We will hold the first Katsura gathering on May 31 at 18:00. The meeting place is the Katsura-River station neighborhood. The content is a workshop for postgraduate examination. Feel free to drop by!

Chapter 5

Probability distribution inference

5.1 HEF extension for probability distribution inference

5.1.1 Optimal association building

We make associations between topics and aspects based on relevance $R(a, t)$. Our approach assumes that each aspect consists of many topics. Here, since we consider that important topics for each aspect have high relevance, an effective strategy of association building connects topics to aspects based on the strength of the relevance. We arranged the topics in descending order of the relevance strength in each aspect and divided the topics into two sets. Our purpose is to discover a significantly high dividing point between a set of topics with high and low relevance. A set of topics with high relevance is our candidate of associations. To achieve this, we adopt a t value in Welch's t-test [77], which is a certification test between two independent groups. When the Welch's t-test value exceeds a threshold, two independent groups are significantly different.

Topic set T_a in aspect a is given as follows:

$$T_a = \underset{T_x \subset T}{\text{argmax}}\ \text{t-test}(T_x, T_y|a),\ \ T_y = T \setminus T_x, \tag{5.1}$$

where T_x denotes the set of topics with high relevance. T_y denotes the complement set of T_x in all topics T extracted by LDA. T_a is given as T_x when the t-test value between T_x and T_y is the highest in all the dividing points. Welch's t-test is defined as follows:

$$\text{t-test}(T_x, T_y|a) = \frac{\mu_x - \mu_y}{\sqrt{\frac{\sigma_x}{|T_x|} + \frac{\sigma_y}{|T_y|}}}, \tag{5.2}$$

$$\mu_i = \frac{1}{|T_i|}\sum_{t \in T_i} R(a,t),\ \sigma_i = \sqrt{\frac{1}{|T_i|}\sum_{t \in T_i}\{R(a,t) - \mu_i\}^2}, \tag{5.3}$$

where $|T_x|$ and $|T_y|$ denote the number of topics. Normalized both relevances of $\hat{Ra}(a,t)$ and $\hat{Rt}(a,t)$ are calculated as follows:

$$\hat{Ra}(a,t) = \frac{R(a,t)}{\sum_{t' \in T_a} R(a,t')}\ \ ,\ \ \hat{Rt}(a,t) = \frac{R(a,t)}{\sum_{a' \in A} R(a',t)}, \tag{5.4}$$

where T_a denotes the topics associated with aspect a. A denotes all the aspects.

5.1.2 Inference

To infer the probability distribution of real life aspects for unknown tweets, we use aspect scores calculated in Eq. (3.8). Aspect a probability $p(a|tw)$ for tweet tw is calculated as follows:

$$p(a|tw) = \frac{Score(a, tw)}{\sum_{a' \in A} Score(a', tw)}. \tag{5.5}$$

5.2 Experimental evaluations

To clarify the effectiveness of our proposed method that infers the probability distribution, we evaluated the JS divergence (JSD) and the Euclidean distance (ED) between each method's inferred and correct probability distributions. As baseline methods, we adopted Naive Bayes, SVM, and L-LDA. To extract topics by LDA and evaluate estimation performance of each method, we used Tweets explained in **Section 4.2.1** and **4.2.1**.

5.2.1 Dataset

Single label dataset for training

To identify the most appropriate aspect for each tweet, we extracted the aspect selected as the top candidate assigned by two or three examinees. The number of labels of each aspect is shown in the "single label" column in **Table 5.1**. The Eating aspect received the most labels: 136 out of 1,500 tweets. Eight aspects were labeled by 100 tweets. The total number of aspects labeled by tweets was 1,345. The tweets, which didn't completely match by the three examinees, was 155 ($= 1,500 - 1,345$).

Multi-label dataset for training

The appropriate several aspects for each tweet is given at least once a selected as the 1st candidate aspects from either three examinees. Therefore, multi-label dataset is superset of single label dataset. The number of labels of each aspect is shown in "multi-label" column in **Table 5.1**. The aspect of Eating and Event are the most labeling ones. The aspects of Contact, Event, Expense, and Locality increased over twice number of labels compared with single label dataset.

Probability distribution dataset for evaluation

To give the probability distribution of the aspects for each tweet, we used all the candidate aspects assigned by the three examinees. Based on the reciprocal rank (RR) [72], which is one evaluation metric for search engine effectiveness, we assumed that the aspects selected with a higher rank have greater weight for the tweets. Correct probability distribution $P(a|tw)$ of each aspect a in tweet tw is shown as follows:

$$RR(a|tw) = \frac{1}{|A|} + \sum_{e \in E} \frac{1}{rank(a|tw, e)}, \tag{5.6}$$

$$P(a|tw) = \frac{RR(a|tw)}{\sum_{a' \in A} RR(a'|tw)}, \tag{5.7}$$

where E denotes all the examinees. $rank(a|tw, e)$ is a candidate number: 1st, 2nd, and 3rd rankings of aspects a labeled by examinee e for tweet tw. The $\frac{1}{|A|}$ is a constant value for probability distribution smoothing. In this paper, we gave by the reciprocal value is given by the aspect number. Probability $P(a|tw)$ of aspect a is given as the value divided by the summation of RR.

Parameter settings

LDA requires hyper parameters. Based on related works[19], we set α to $\frac{50}{|T|}$ and β to 0.1. $|T|$ denotes the number of topics chosen from among 50, 100, 200, 500, and 1,000 topics in **Section 5.2.4**. The iterative calculation count in LDA is 100 times in every case.

5.2.2 Evaluation metrics

To correctly evaluate our method performance, we used 10-fold cross validation. We evaluate the JSD and ED between the inference and correct probability distributions. JSD is a metrics that measures the similarity among probability distributions [50]. When both metrics are low, our method accurately infers the probability distribution

Table 5.1: Number and probability of labels by aspect

	Single label			Multi-label		
Aspects	Number	$P(a)$	$\lvert T_a \rvert$	Number	$P(a)$	$\lvert T_a \rvert$
Appearance	104	0.0773	341	151	0.0636	329
Contact	100	0.0743	343	208	0.0877	341
Disaster	39	0.0290	**379**	52	0.0219	363
Eating	136	0.1011	247	219	0.0923	296
Event	85	0.0632	241	219	0.0923	288
Expense	76	0.0565	334	211	0.0889	**387**
Health	92	0.0684	348	121	0.0510	322
Hobby	108	0.0803	339	200	0.0843	312
Living	97	0.0721	332	141	0.0594	328
Locality	68	0.0506	320	147	0.0619	348
School	110	0.0818	321	153	0.0645	323
Traffic	107	0.0796	346	136	0.0573	306
Weather	111	0.0825	291	157	0.0662	291
Working	105	0.0781	299	176	0.0742	303
Other	7	0.0052	248	82	0.0346	375
Total	1,345	1.0000		2,373	1.0000	

of tweets. JSD and ED between the probability distributions of x and y are calculated as follows:

$$\mathrm{JSD}(x, y) = \frac{1}{2}\left(\sum_{a \in A} x(a) \log \frac{x(a)}{z(a)} + \sum_{a \in A} y(a) \log \frac{y(a)}{z(a)}\right), \tag{5.8}$$

$$\mathrm{ED}(x, y) = \sqrt{\sum_{a \in A} \{x(a) - y(a)\}^2}, \tag{5.9}$$

where $z(a)$ denotes the average of $x(a)$ and $y(a)$.

5.2.3 Baseline methods

We extracted nouns, verbs, and adjectives using a Japanese morphological analyzer called MeCab [37] and entered the sets of words and label(s) to every method in common.

Uniform distribution (UD)

As the most simplest comparison method, we prepared the uniform distribution of aspects, each of which has $\frac{1}{|A|}$ probability. $|A|$ is the number of aspects.

Prior distribution (PD)

Prior distribution is calculated from the ratio of the number of aspects in the training dataset. UD and PD do not depend on the set of words appearing in the tweets.

Naive Bayes (NB)

A Naive Bayes classifier [15], which is one of the most typical and effective classification methods, classifies the labels with the highest posterior probability for a document. In our experimental evaluations, we used the normalized posterior probability of each document.

Support vector machine (SVM)

We used LIBSVM [12] as a support vector machine library. LIBSVM provides a probability estimation tool [79] for each class in addition to document classification. As SVM parameters, we chose a linear kernel and set parameter C to 1.0, indicated by a grid search in the LIBSVM tools [24].

Labeled LDA (L-LDA)

Labeled LDA is an LDA extended models, which was proposed by Ramage et al. [58]. L-LDA sets the hyperparameters of both α and β, as in LDA. We experimentally set

α to 0.1 and β to 0.1, and the iterative calculation count in L-LDA was 100.

5.2.4 Experimental results

Comparison of number of topics

We evaluated the micro-average value of JSD between the inference and correct probability distributions in both the single and multi-label cases (**Table 5.2**). In both cases, according to an increasing number of topics, JSD decreased. Its decrease stabilized from 500 topics because the JSD difference at 500 and 1,000 topics is slight. A minimum JSD was achieved at 1,000 topics in both the single and multi-label cases. Therefore, we used 1,000 as the optimal number of topics for HEF.

Number of topics connected to each aspect

We show the number of topics associated to each aspect in the $|T_a|$ column of **Table 5.1**. These numbers are optimized by Welch's t-test. The maximum topic numbers of single and multi-label cases are the aspects of Disaster at 379 and Expense at 387 respectively. The minimum topic number is the Event aspect in both the single and multi-label cases.

We show the relevance and the t-test distributions of the Disaster and Event aspects in **Figs. 5.1** and **5.2**. The horizontal axes of all the figures are the topic rankings that are arranged in descending order of the relevance strength. The left and right vertical axes of both figures are the relevance and t-test values. The Disaster aspect achieved the maximum t-test value from 300 to 400 topics. On the other hand, the Event aspect was achieved the maximum from 200 to 300 topics.

We show the relevance and the t-test distributions of each aspect in **Fig. 5.3** and **5.4**. The horizontal axes of both figures are the topic rankings that are arranged in descending order of the relevance strength. In each figure, the strength of the value is shown by a color chart mapped from red to white. For example, in **Fig. 5.3**, the color of every aspect changes from red to white based on the decreased topic ranking

because topics are arranged in descending order of relevance strength. The Disaster aspect more rapidly changed from red to blue compared to other aspects.

In **Fig. 5.4**, every aspect was red from 200 to 500 topics, suggesting that every aspect is optimally divided in this range. Here, the topics with the highest t-test value in each aspect are shown in the $|T_a|$ column of **Table 5.1**. The Disaster aspect is nor red or yellow because its relevance strength is low.

Comparison of association building methods

To evaluate the effectiveness of our association building method, we implemented three simple methods to build associations; first, we associated a topic with the highest relevance to each aspect; second, we associated ten topics with higher relevance to each aspect; finally, we associated all topics to each aspect.

The JSD value by each method is shown in **Table 5.3**. The minimum JSD value was achieved by t-test topics. The first and second methods showed higher JSD values than the third method. Based on these results, the aspect architecture is insufficient in for just a few topics. However, to build refined associations, the architecture needs to delete extra topics from the third method's result.

Inference performance of each method

We show the micro-average value and the standard deviation of JSD and ED by each method in **Figs. 5.5** and **5.6**. The vertical axis shows the JSD and ED values. We took a one-sided t-test of HEF's JSD and ED values against the baseline methods' values. That result was drawn on the top of each baseline method as "*" symbols in the figures; "***" represents a significantly-high value at 1%, "**" at 5%, and "*" at 10%.

From the t-test results, our method efficiently estimated the probability distributions against all the baseline methods in the single label case. In the multi-label case, HEF performed significantly better than every baseline method except SVM.

Table 5.2: JSD scores in each number of topics in HEF

Number of topics	Single label	Multi-label
50	0.2408	0.2170
100	0.2324	0.2127
200	0.2159	0.1977
500	0.1987	0.1852
1,000	**0.1926**	**0.1820**

Table 5.3: JSD by each association building method

Method	Single label	Multi-label
Highest topic	0.3376	0.2921
Highest 10 topics	0.2391	0.2281
All topics	0.2068	0.1935
t-test topics	**0.1926**	**0.1820**

5.3 Discussions

From **Figs. 5.5** and **5.6**, the multi-label dataset achieved lower JSD and ED values than the single label dataset, except for the uniform distribution. This reason is clear because multi-label dataset has more detailed training information than the single label one to infer the probability distributions. SVM especially decreased the JSD values of 0.04 ($= 0.22-0.18$) in the multi-label case compared with the single label one. These results suggest that SVM outperformed HEF when it built a model by training datasets with more detailed information, such as the multi-label classification dataset used in Chapter 4. On the other hand, generally, a cost for obtaining multi-label training datasets by human annotation is more than single-label ones. Therefore, we think important to estimate the correct probability distribution of aspects by single-label datasets.

Our method showed the lowest JSD and ED values in both the single and multi-label cases. In the single label case, HEF showed significantly higher performance than the other methods. We can see an optimal example tweet that explains this reason in **Table 5.4** and **Fig. 5.7**. **Table 5.4** shows the example tweet sentence and its labels. The main topic of this tweet is open campus, and two examinees selected the School aspect as its top candidate. Therefore, this tweet received the School aspect label. On the other hand, examinee E3 selected the Event aspect as its top candidate because he defined open campus as an event. In fact, examinee E1 selected the Event aspect as his second highest candidate. In multi-label cases, this tweet was labeled by School and Event aspects.

Fig. 5.7 shows the correct probability distributions of **Table 5.4**'s tweet as a black solid line. The School and Event aspects have higher probability than the other examinee labeling results. In addition to the same figure, we show the probability distributions estimated by each method that was trained by a single label dataset. We focus on the probabilities of the Event and School aspects. The inferred probability of the School aspect by each method is higher than the other aspects. NB inferred

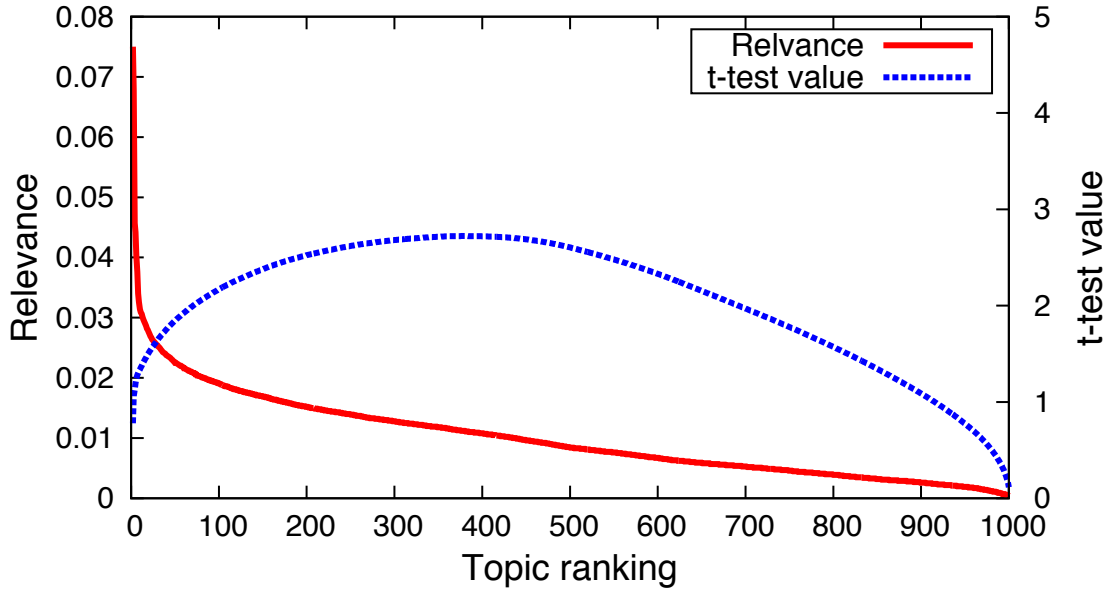

Figure 5.1: Relevance and t-test value distributions of Disaster

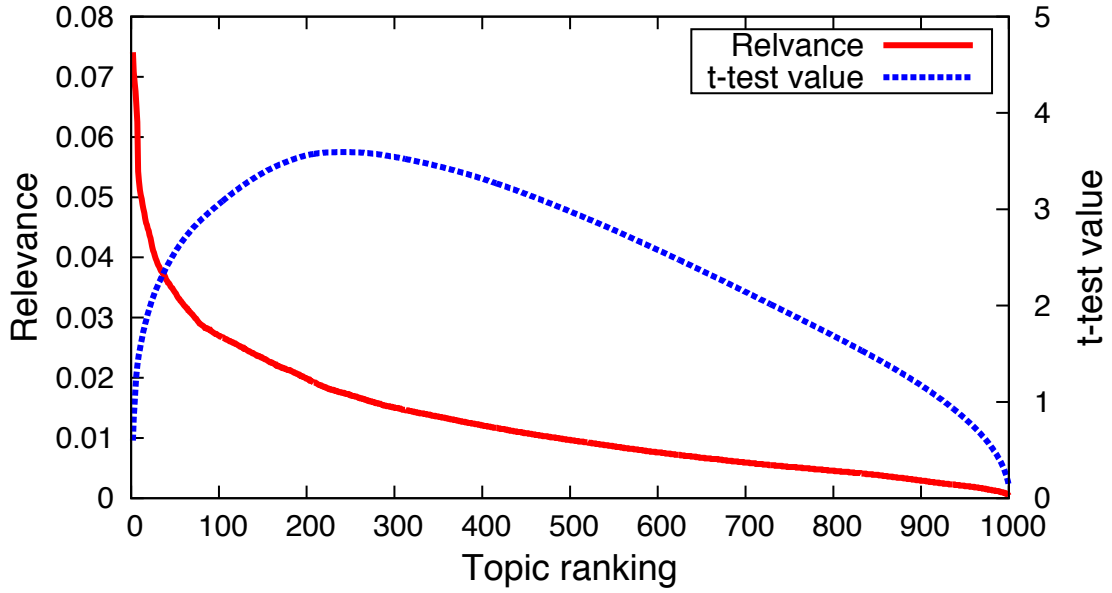

Figure 5.2: Relevance and t-test value distributions of Event

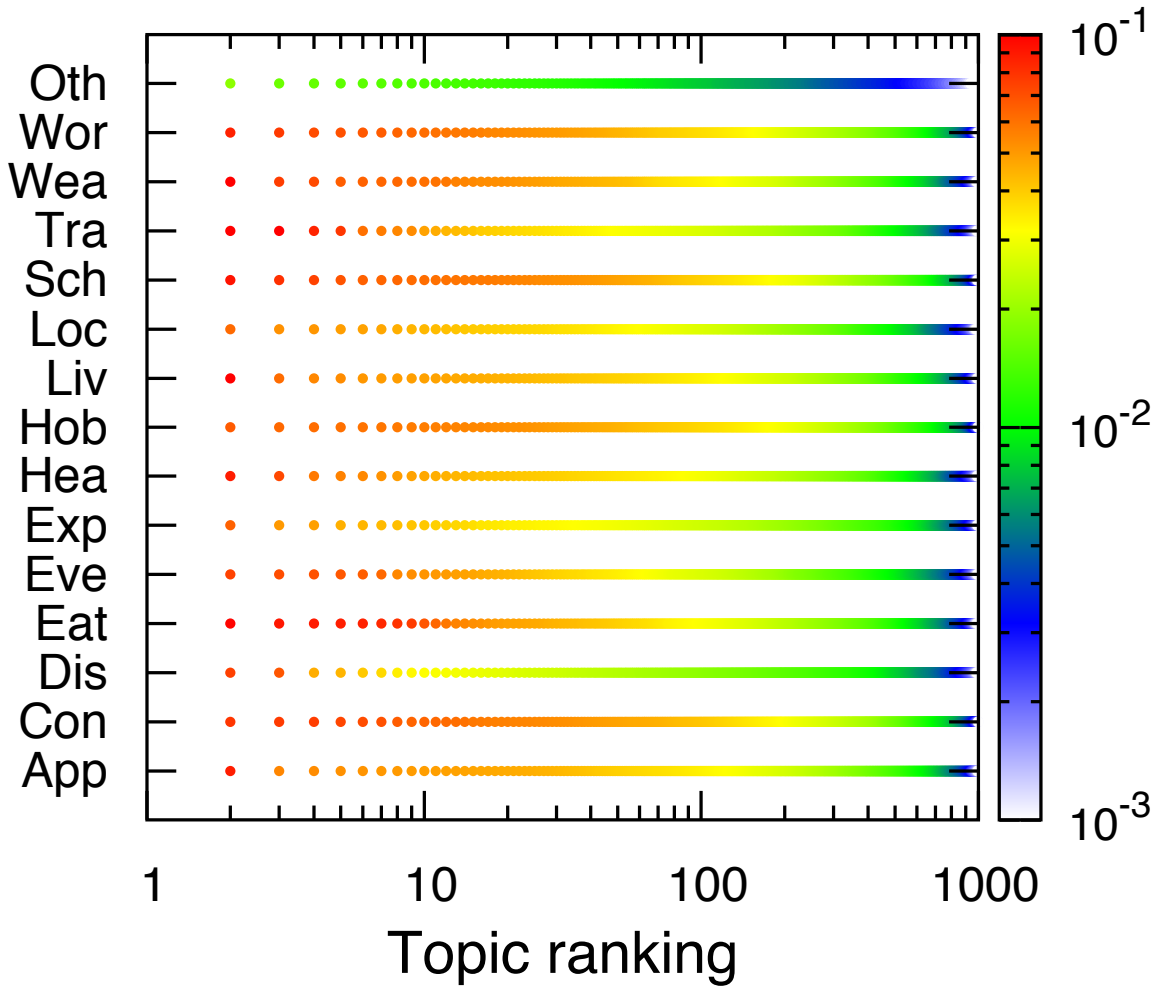

Figure 5.3: Relevance distributions of all aspects

a higher probability than 0.50. SVM and HEF showed a lower probability than the correct one. Next, in the inferred probabilities of Event by each method, HEF successfully estimated the most approximate probability for answering it.

This tweet includes many terms that suggest the School aspect: "university", "professor", and "lecture". However, the only term for estimating the Event aspect is the verb "held". For these reasons, the estimations of NB, SVM, and L-LDA, all of which directly calculate the likelihood of terms, were not appropriate.

Here, we show the high occurrence probability words in the topic associated with the highest relevance to the Event aspect by HEF in **Table 5.5**. This topic includes terms related to the Event aspect: "participation", "held", and "conference". On the

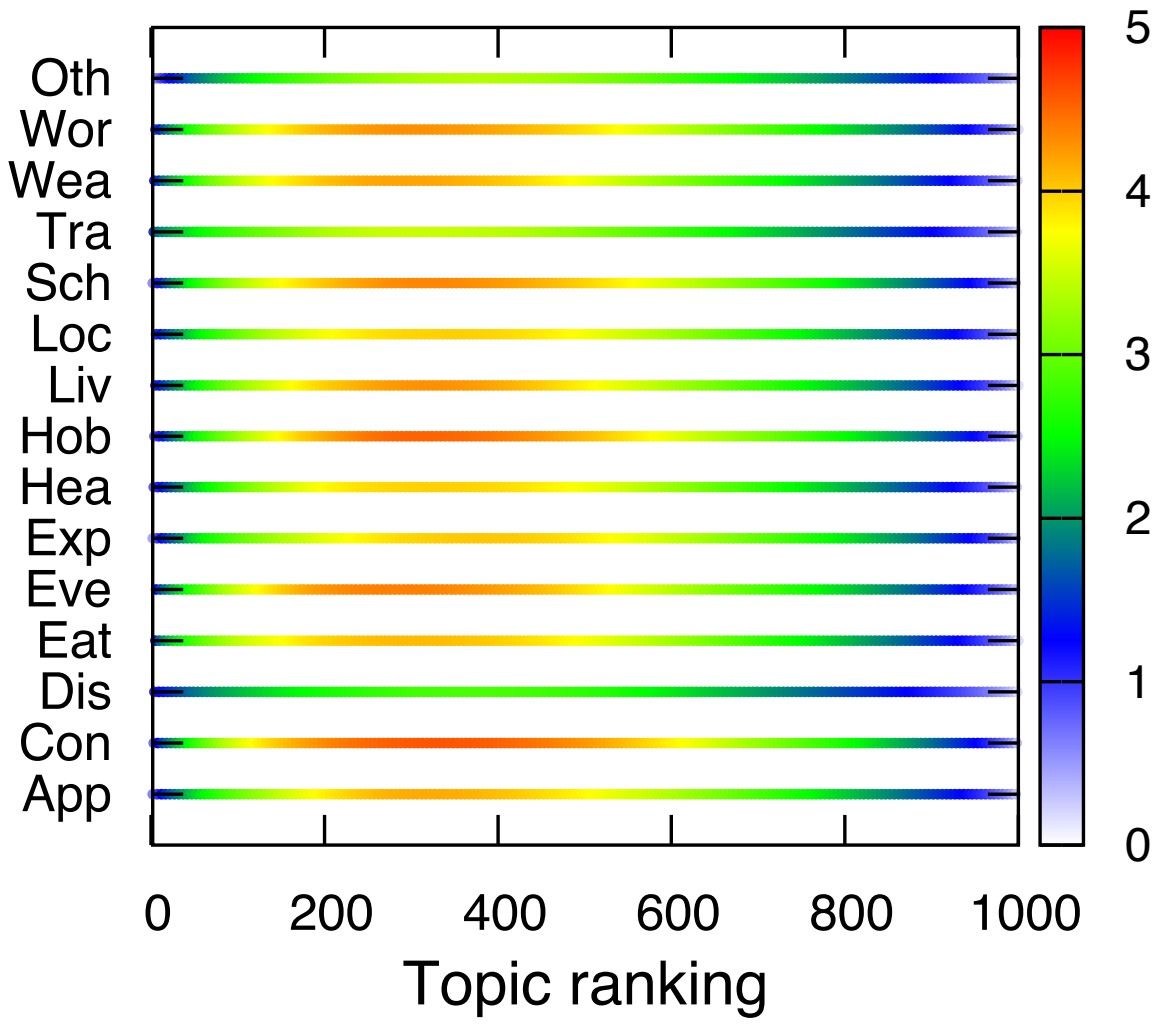

Figure 5.4: t-test value distributions of all aspects

other hand, such terms as “lecture”, “campus”, and “university” are also included in the topic, suggesting that they often appear together in many tweets. Therefore, such terms such as “campus” and “university” are frequently mentioned in connection with Event aspect terms, including “held”. Our method can build associations between this topic and the Event and School aspects because it can use such terms as “held” assigned Event aspect and “lecture” assigned School aspect. So, HEF inferred the Event aspect with high probability for **Table 5.4**’s tweet.

Finally, we show the number of average labels (Mean), its standard deviation (SD), and the assigned labels for a tweet by each examinee in **Table 5.6**. The mean and

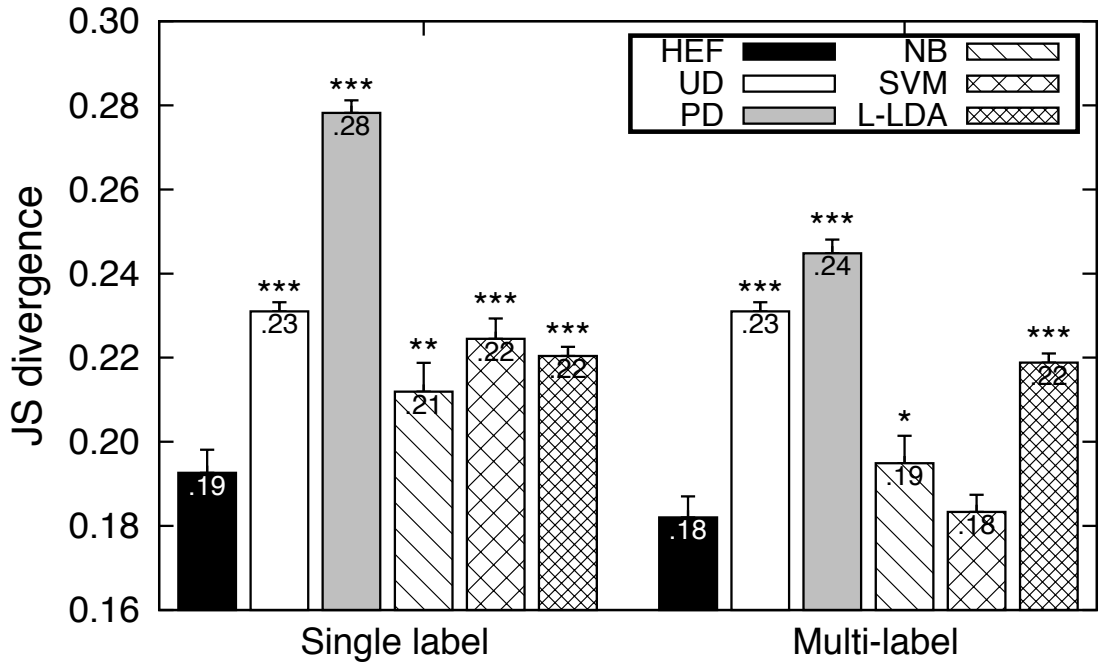

Figure 5.5: JS divergence

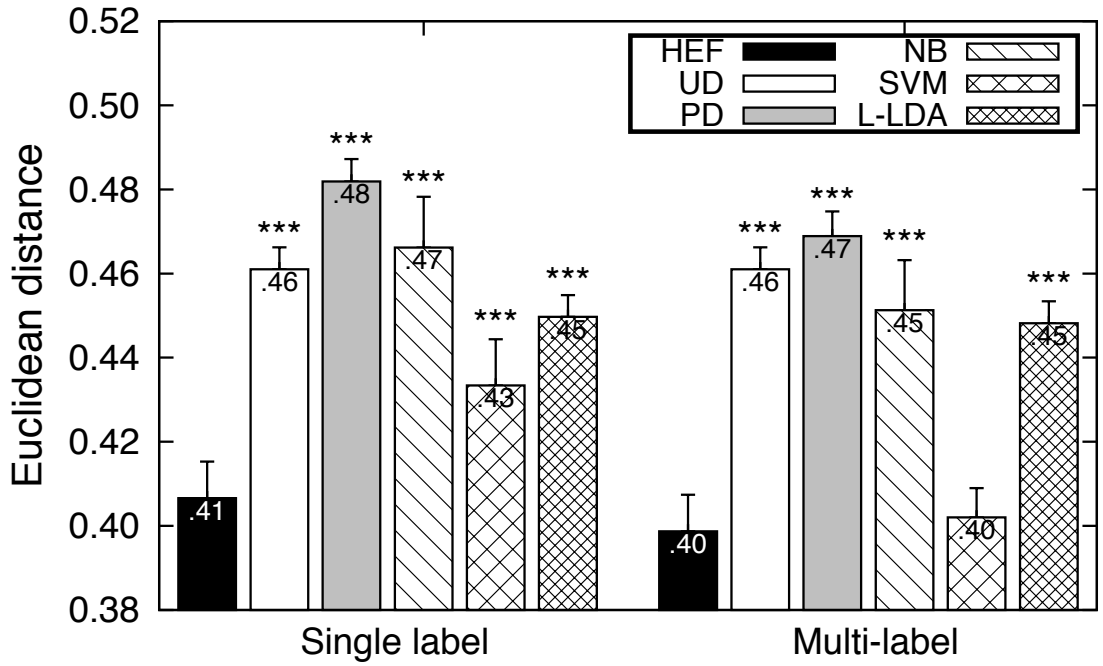

Figure 5.6: Euclidean distance

Table 5.4: Effectively inferred probability distributions of aspects by HEF

Examinees	1st	2nd	3rd
E1	School	Event	Hobby
E2	School	Other	
E3	Event	School	Other
Tweet	We'll hold an open campus for Kyoto Seika University on June 9, and some professors will provide special lectures!		
Single label	School		
Multi-label	School, Event		

Table 5.5: High occurrence probability terms in highest relevance topic associated to Event

Topics	Characteristic words
Topic 387	participation, Kyoto, lecture, held, hall, culture, campus, conference, university

standard deviation of examinees E1 and E3 are approximate values. The number of assigned labels for them is also shown as similar distributions. However, the number of average assigned labels by examinee E2 shows greater values than E1 and E3. E2 tended to assign many labels for a tweet from the values in the Three labels column. These results suggest that the criteria of the users for requiring aspects are different. For example, E1 and E2 are more accuracy-oriented users and E3 is an exhaustive-oriented user. A multi-label classification approach has difficulty accommodating an individual user's requirements or various situations. However, the representation of probability distribution on tweets can be applied to these users.

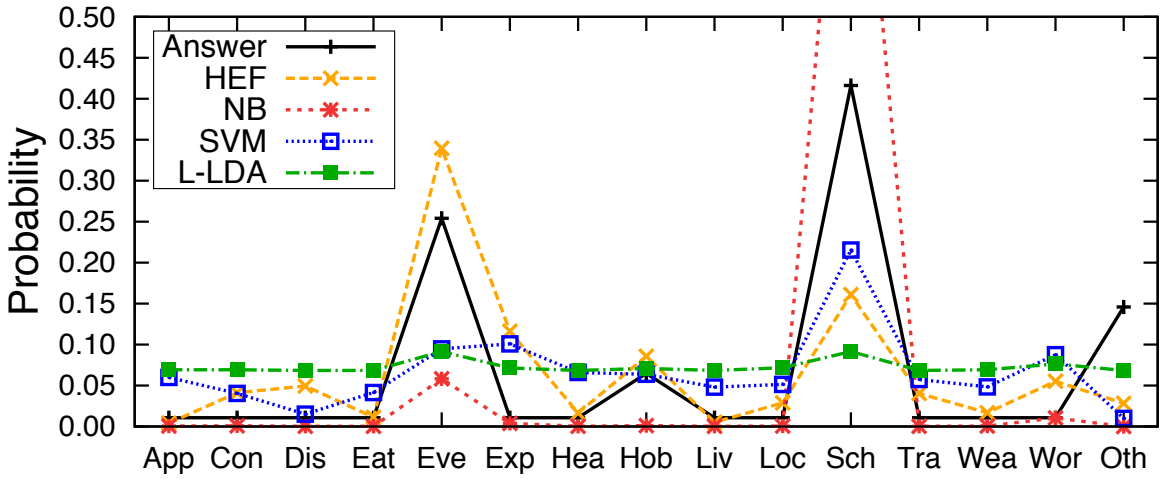

Figure 5.7: Probability distributions of aspects estimated by each method for **Table 5.4**'s tweet

Table 5.6: Number of average labels for a tweet in each examinee

	Mean	SD	One label	Two labels	Three labels	Total
E1	1.519	0.633	834 (55.6%)	553 (36.9%)	113 (7.5%)	1,500
E2	2.498	0.700	180 (12.0%)	393 (26.2%)	927 (61.8%)	1,500
E3	1.497	0.625	859 (57.3%)	536 (35.7%)	105 (7.0%)	1,500

Chapter 6

Discussions

6.1 Achievements in this dissertation

This section summarizes the achievements of the four challenges explained in Chapter 1.

Achievement 1 **Table 4.7** nicely represents the effectiveness of HEF for short sentence estimation. Its maximum average value was 5.15 for the number of labelings in all comparison methods. In other words, HEF approximately estimated five aspects for each tweet whose maximum number of characters is 140. If it randomly estimates five aspects of each tweet, the precision value will be the lowest in all the methods because the number of labelings of other methods is lower than HEF. However, HEF's precision value is second highest at 0.63 after NBML. This suggests that HEF frequently estimates extra aspects in addition to the correct labels of each tweet. That example was shown in **Table 4.13**, which was estimated an extra Locality aspect by HEF. From these results, HEF achieved appropriate estimation for short sentences using tweets as demonstrations.

Achievement 2 In multi-label classification, though the precision, recall, and f-measure values of the typical methods rapidly dropped when the training data

decreased, HEF retained its high evaluation values in **Figs. 4.19**, **4.20**, and **4.21**. In the probability distribution inference, HEF showed significantly lower JSD and ED values than the typical methods when it built associations between the topics and the aspects by the single label dataset in **Fig. 5.5** and **5.6**. As reasons why HEF appropriately estimated the aspects of unknown tweets, we believe that HEF effectively expanded the terms using topics extracted by LDA and calculated the aspect scores from unknown terms that do not appear in the training dataset. Moreover, we confirmed that the number of topics connected to the aspects was adjustd by the precision and recall curves in **Fig. 4.1**. Therefore, HEF can obtain higher performance by selecting the optimal number of topics for the given tasks. These results suggest that HEF achieved high estimation performance using a small amount of training data, as described in Challenge 2.

Achievement 3 Chapter 4 demonstrated that our proposed method appropriately estimates several aspects for unknown tweets. In the comparison evaluations shown in **Table 4.4**, **4.5**, and **4.6**, the fundamental proposal method (HEF0) achieved the highest recall value at 0.67 in the baseline methods. HEF, which introduced an entropy feedback mechanism, showed the maximum f-measure value at 0.63 because it was the refined associations between the topics and the aspects. From these results, HEF achieved multi-label classification which estimated several aspects of unknown tweets.

Achievement 4 Chapter 5 showed that HEF effectively infers the aspect probability distribution of unknown tweets. Our sophisticated experimental evaluations clarified that our proposed method achieved significantly higher estimation performance than typical methods from **Figs. 5.5** and **5.6**. Although the JSD and ED values of SVM in single label training greatly increased more than multi-label training, HEF retained low values in both situations. From these results, HEF appropriately inferred the probability distribution of unknown tweets by

training using labeled data.

6.2 Associations with latent topics

This section summarizes the advantages and weaknesses of the associations between the topics and the aspects based on the experimental evaluations results of both Chapters 4 and 5. First, Chapter 4 showed that HEF0 achieved maximum recall and minimum precision in all the comparison methods. The basic purpose for making associations with latent topics is to expand the terms using topics and to increase the completeness in the estimation aspects of unknown tweets. Although HEF0 achieved it purpose, terms were excessively expanded because the fundamental associations with the topics are competitive among the aspects. For example, when regional names appeared in unknown tweets, HEF0 estimated many aspects such as Disaster, Event, Locality, and Traffic in **Table 4.12**'s tweet since these aspects are strongly associated to the same topics. Therefore, HEF0 had many wrong estimations, which lowered the precision.

To overcome this weakness, HEF was introduced into the entropy feedback mechanism, which can refine competitive topics among aspects. Although HEF's recall decreased by 0.04 $(0.63 - 0.67)$, it's precision increased by 0.16 $(0.63 - 0.47)$ and HEF achieved the highest F-measure among all of the compared methods. In other words, HFE overcame the weakness by the entropy feedback mechanism. Even if the number of aspects increases, we believe that HEF shows higher estimation performance than the other methods because it is an estimation model that isn't dependent on the number of aspects.

HEF can incorporate other topic models and matrix factorization algorithms as well as LDA because in the first phase, HEF essentially extracts the latent topics from documents. Each topic has similar terms with high occurrence probability in the documents. Even if the topic tendency differs among topic models and matrix

factorization algorithms, HEF can enhance the estimation performance because it associates several feature topics to given human labels, such as life aspects, by entropy feedback mechanism and the optimal association building method. In the future, we will evaluate the HEF performance by extracting topics using various topic models.

On the other hand, the associations between the topics and the aspects have to decide some parameters such as the number of topics and d in Eq. (3.7). The optimal number of topics achieved by the maximum F-measure was obtained by maximizing JS_{sum}. Although the optimal number of topics to achieve the maximum F-measure was obtained by maximizing JS_{sum}, optimal parameter d must explore the value that maximizes the F-measure in each aspect. To optimize the associations between the topics and the aspects, Chapter 5 proposed an association building method based on the t-test. It achieved the minimum JSD in all the association building strategies in both cases of single label and multi-label training. However, the number of topics is not automatically optimized in it. The optimizing all the parameters is future work.

Finally, we discuss the single label estimation task, which classifies the most suitable label for each datum. Chapter 5's results suggest that HEF is less effective in this task compared with multi-label classification and probability distribution inference tasks. The correct single aspect in **Table 5.4's** tweet is School because the two examinees selected it as the most suitable aspect for this tweet. From **Fig. 5.7**, which shows the probability distributions of the aspects estimated by the comparison methods for that tweet, Event is the aspect with the highest probability by HEF. Generally, the single label estimation task gives a label with the highest probability in all the labels. Therefore, HEF estimation is wrong in this case. The single label estimation task is important to obtain a clearly higher score than the other aspects, but HEF does not satisfy that requirement based on these dissertation evaluation results. However, we believe that the number of life aspects, which are imagined from a tweet by human, is essentially multi-labels. We think important to achieve the mild estimation like HEF to infer several life aspects of each tweet.

Chapter 7

Conclusion

This dissertation proposed the life aspect inference method by hierarchical estimation framework (HEF) based on associations between topics and aspects. This framework feature is composed of two phase semi-supervised machine learning, in which many topics are extracted from a sea of tweets using an unsupervised machine learning model LDA. Associations among many topics and fewer aspects are built using labeled tweets. Using topics, aspects are associated with various keywords by a small set of labeled tweets.

Chapter 4 extended HEF to multi-label classification to clearly provide real life information with particular aspects for users. To refine the associations between topics and aspects, HEF introduced the entropy feedback mechanism, which iteratively calculates feedback coefficients calculated by entropy between topics and aspects. In experimental evaluations using actual collected real life tweets, our prototype system demonstrated that HEF can appropriately estimate some aspects of all the unknown tweets. Entropy feedback effectively refines the associations between topics and aspects. HEF showed the highest F-measure among typical methods of multi-label classification. With less training data, the precision, recall, and F-measure values of the typical methods rapidly dropped; however, HEF retained its high evaluation values. Especially in F-measure, HEF usually achieved the highest score in every

method.

Chapter 5 extended HEF to life aspect distribution inference to provide real life information on specific aspects according to user orientation such as users with exhaustive and accuracy oriented. HEF introduced the optimal association build method based on t-test, which is an efficient strategy to manage the relationships between topics and aspects. This study challenge is to train from labeled tweets and infer the probability distribution of the aspects of unknown tweets based on a natural extension of HEF. The experimental evaluations of this study prepared a small set of labeled tweets based on the classifications of three examinees and calculated probability distributions of each tweet from them. In the case of single label training, HEF showed significantly lower JS Divergence and Euclidean Distance values than every baseline method based on sharing topics by several aspects.

From Chapter 4 and 5 results, HEF scheme is an effective life aspect inference methods of the multi-label classification and probability distribution using a small labeled dataset for such short sentences as tweets because associations between topics and aspects appropriately expanded terms. In the future, we will confirm the effectiveness of our method using other datasets, such as newspapers and blogs.

Acknowledgements

I sincerely thank to my supervisor, Professor Tetsuji Satoh, who graciously supported and guided me for five years while I worked on this doctoral dissertation.

I am grateful to my sub-supervisors, Professor Atsuyuki Morishima and Associate Professor Taro Tezuka, both of whom shared their time and knowledge with me.

I thank Professors Atsushi Toshimori and Hiroyuki Kitagawa who provided many essential and beneficial comments their reviews of my dissertation.

I would like to thank to Professor Noriko Kando, National Institute of Informatics, and Associate Professor Hideo Joho, both of whom gave the deep considerations from information retrieval philosophy to my dissertation.

I also thank all the students in my laboratory with whom I shared many pleasures and setbacks. Discussions with Tetsuro Ohyama, Yutaro Yamaguchi, Yuki Doumae, and Yoshiki Nakaoka were especially insightful.

Finally, I express my deepest thanks to my family and my friends for their infinite encouragement and kindness.

This study was supported by Grant-in-Aids for scientific research No.25280110 and No.15J05599 and by NII's strategic open-type collaborative research. For participation at international academic conferences, this study's travel expenses and registration fees were subsidized by the Hara Research Foundation, the Telecommunication Advancement Foundation, and the educational strategic expense of the Faculty of Library, Information and Media Studies, University of Tsukuba.

Bibliography

[1] 2011 tohoku earthquake and tsunami. https://en.wikipedia.org/wiki/2011_Tohoku_earthquake_and_tsunami.

[2] Twitter. https://twitter.com.

[3] Twitter search api. https://dev.twitter.com/docs/api/1/get/search.

[4] Yutaka Arakawa, Shigeaki Tagashira, and Akira Fukuda. Relationship analysis between user ' s contexts and real inputwords through twitter. In *Proceedings of the IEEE Globecom 2010 Workshop on Ubiquitous Computing and Networks(UbiCoNet 2010)*, pages 1813–1817, 2010.

[5] Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. Twitter catches the flu: Detecting influenza epidemics using twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1568–1576, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[6] Junki Arimitsu, Qiang Ma, and Masatoshi Yosikawa. A user experience-oriented microblog retrieval method. In *The 3rd Forum on Data Engineering and Information Management*, F5-2, 2011 (in Japanese).

[7] David M. Blei and John D. Lafferty. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 113–120, New York, NY, USA, 2006. ACM.

[8] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[9] Johan Bollen, Alberto Pepe, and Huina Mao. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *CoRR*, abs/0911.1583, 2009.

[10] Meeyoung Cha, Hamed Haddadi, Fabrcio Benevenuto, and Krishna P. Gummadi. Measuring user influence in twitter: The million follower fallacy. In *Proceedings of international AAAI Conference on Weblogs and Social*, pages 10–17, 2010.

[11] Wen Chan, Jintao Du, Weidong Yang, Jinhui Tang, and Xiangdong Zhou. Term selection and result reranking for question retrieval by exploiting hierarchical classification. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 141–150, New York, NY, USA, 2014. ACM.

[12] Chihchung Chang and Chihjen Lin. Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):1–27, May 2011.

[13] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. You are where you tweet: A content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 759–768, New York, NY, USA, 2010. ACM.

[14] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

[15] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Journal of Machine Learning Research*, 20(3):273–297, September 1995.

[16] Qiming Diao, Jing Jiang, Feida Zhu, and Ee-Peng Lim. Finding bursty topics from microblogs. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 536–544, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[17] Pedro Domingos and Michael Pazzani. On the optimality of the simple bayesian classifier under zero-one loss. *Journal of Machine Learning Research*, 29(2-3):103–130, 1997.

[18] Yuki Doumae and Yohei Seki. Estimation of twitter user's life-area using area related terms selected by semi-supervised topic model. *Journal of IPSJ:Transactions on Databases (TOD)*, 7(3):1–13, 2014.

[19] Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. *National Academy of Sciences of the United States of America*, 101:5228–5235, 2004.

[20] John Hannon, Mike Bennett, and Barry Smyth. Recommending twitter users to follow using content and collaborative filtering approaches. In *Proceedings of the Fourth ACM Conference on Recommender Systems*, RecSys '10, pages 199–206, New York, NY, USA, 2010. ACM.

[21] Yuki Hattori and Akiyo Nadamoto. Extracting tip information from social media. In *Proceedings of the 14th International Conference on Information Integration and Web-based Applications and Services*, pages 205–212, 2012.

[22] Charmgil Hong, Iyad Batal, and Milos Hauskrecht. A mixtures-of-experts framework for multi-label classification. *CoRR*, abs/1409.4698, 2014.

[23] Liangjie Hong and Brian D. Davison. Empirical study of topic modeling in twitter. In *In proceedings of the First Workshop on Social Media Analytics*, pages 80–88, 2010.

[24] Chihwei Hsu, Chihchung Chang, and Chihjen Lin. A practical guide to support vector classification. `http://www.csie.ntu.edu.tw/\~cjlin/papers/guide/guide.pdf`, 2010.

[25] Sha Hu, Zhicheng Dou, Xiaojie Wang, Tetsuya Sakai, and Ji-Rong Wen. Search result diversification based on hierarchical intents. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, pages 63–72, New York, NY, USA, 2015. ACM.

[26] Kayo Ikeda, Katsuyoshi Tanabe, Hidenori Okuda, and Masahiro Oku. A web-mining technique of experience information. *Transactions of Information Processing Society of Japan*, 49(2):838–847, 2008.

[27] Kentaro Inui, Shuya Abe, Kazuo Hara, Hiraku Morita, Chitose Sao, Megumi Eguchi, Asuka Sumida, Koji Murakami, and Suguru Matsuyoshi. Experience mining: Building a large-scale database of personal experiences and opinions from web documents. In *Web Intelligence*, pages 314–321, 2008.

[28] Aya Ishino, Hidetsugu Nanba, and Toshiyuki Takezawa. Automatic compilation of an online travel portal from automatically extracted travel blog entries. In Rob Law, Matthias Fuchs, and Francesco Ricci, editors, *Information and Communication Technologies in Tourism 2011*, pages 113–124. Springer Vienna, 2011.

[29] Aya Ishino, Hidetsugu Nanba, and Toshiyuki Takezawa. Extracting transportation information and traffic problems from tweets during a disaster: Where do you evacuate to? In *The Second International Conference on Advances in Information Mining and Management (IMMM2012)*, 2012.

[30] Yusuke Iwaki, Adam Jatowt, and Katsumi Tanaka. Supporing finding read-valuable articles in micro-blogs. In *The 1st Forum on Data Engineering and Information Management*, A6-6, 2010 (in Japanese).

[31] Yuichiro Kase and Takao Miura. Mining classes by multi-label classification. In *15èmes Journées Francophones Extraction et Gestion des Connaissances, EGC 2015, 27-30 Janvier 2015, Luxembourg*, pages 77–82, 2015.

[32] Hiroki Kawaguchi, Shinpei Matsumoto, , and Fujio Toriumi. A method to quantify twitter user's posting activities for constructing disaster information support system. In *Proceedings of the 2014 IEEE 7th International Workshop on Computational Intelligence and Applications (IWCIA)*, pages 137–140, 2014.

[33] Hideto Kazawa, Tomonori Izumitani, Hirotoshi Taira, and Eisaku Maeda. Maximal margin labeling for multi-topic text categorization. In *Advances in Neural Information Processing Systems 17*, pages 649–656. MIT Press, 2005.

[34] Dongwoo Kim, Suin Kim, and Alice Oh. Dirichlet process with mixed random measures: A nonparametric topic model for labeled data. In John Langford and Joelle Pineau, editors, *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 727–734, New York, NY, USA, 2012. ACM.

[35] Tasuku Kimura and Hisashi Miyamori. A method of classifying relationships between hashtags using co-occurrence and latent topics. *IEICE TRANSACTIONS on Fundamentals of Electronics, Communications and Computer Sciences*, J98-D(8):1151–1161, 2015.

[36] Daichi Koike, Yusuke Takahashi, Takehito Utsuro, Masaharu Yoshioka, and Noriko Kando. Time series topic modeling and bursty topic detection of correlated news and twitter. In *Proceedings of the 6th International Joint Conference on Natural Language Processing*, pages 917–921, 2013.

[37] Taku Kudo. Yet another part-of-speech and morphological analyzer. `http://mecab.sourceforge.net/`, 2005.

[38] Takeshi Kurashima, Tomoharu iwata, Takahide Hoshide, Noriko Takaya, and KO Fujimura. Joint modeling of user's activity area and interests for local rec-

ommendation. *Journal of IPSJ:Transactions on Databases (TOD)*, 6(2):30–41, 2013.

[39] Takeshi Kurashima, Taro Tezuka, and Katsumi Tanaka. Blog map of experiences: Extracting and geographically mapping visitor experiences from urban blogs. In *Proceedings of the 6th International Conference on Web Information Systems Engineering*, pages 496–503, 2005.

[40] Ryong Lee, Shoko Wakamiya, and Kazutoshi Sumiya. Urban area characterization based on crowd behavioral lifelogs over twitter. *Personal Ubiquitous Comput.*, 17(4):605–620, April 2013.

[41] Chenliang Li, Aixin Sun, and Anwitaman Datta. Twevent: Segment-based event detection from tweets. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 155–164, New York, NY, USA, 2012. ACM.

[42] Chenliang Li, Jianshu Weng, Qi He, Yuxia Yao, Anwitaman Datta, Aixin Sun, and Bu-Sung Lee. Twiner: Named entity recognition in targeted twitter stream. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 721–730, 2012.

[43] Qiang Ma and Katsumi Tanaka. Retrieving regional information from web by contents localness and user location. In SungHyon Myaeng, Ming Zhou, Kam-Fai Wong, and Hong-Jiang Zhang, editors, *Information Retrieval Technology*, volume 3411 of *Lecture Notes in Computer Science*, pages 301–312. Springer Berlin Heidelberg, 2005.

[44] Zongyang Ma, Aixin Sun, Quan Yuan, and Gao Cong. Tagging your tweets: A probabilistic modeling of hashtag annotation in twitter. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 999–1008, 2014.

[45] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schutze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[46] Michael Mathioudakis and Nick Koudas. Twittermonitor: Trend detection over the twitter stream. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*, SIGMOD '10, pages 1155–1158, New York, NY, USA, 2010. ACM.

[47] Matthew Michelson and Sofus A. Macskassy. Discovering users' topics of interest on twitter: A first look. In *Proceedings of the Fourth Workshop on Analytics for Noisy Unstructured Text Data*, AND '10, pages 73–80, New York, NY, USA, 2010. ACM.

[48] Rada Mihalcea and Paul Tarau. Textrank: Bringing order into texts. In Dekang Lin and Dekai Wu, editors, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 404–411, July 2004.

[49] Yuhiro Mizunuma, Shuhei Yamamoto, Yutaro Yamaguchi, Atsushi Ikeuchi, Tetsuji Satoh, and Satoshi Shimada. Twitter bursts: Analysis of their occurrences and classifications. In *Proceedings of the Eight International Conference on Digital Society (ICDS2015)*, pages 182–187, 2014.

[50] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*, page 58. The MIT Press, 2012.

[51] Shinichi Nagano, Koji Ueno, and Kenta Cho. Development of a system detecting train status information from social sensors. *IEICE TRANSACTIONS on Fundamentals of Electronics, Communications and Computer Sciences*, J96-D(10):2262–2273, 2013 (in Japanese).

[52] Yuto Nakajima, Hirotaka Niitsuma, and Manabu Ohta. Travel route recommendation using tweets with location information. *IPSJ SIG Technical Report*, 2013-DBS-158(28):1–6, 2013 (in Japanese).

[53] Yoko Nishihara, Keita Sato, and Wataru Sunayama. Personal experience acquisition support from blogs using event - depicting images. *Journal of Japan Society for Fuzzy Theory and Intelligent Informatics*, 20(5):757–767, 2008.

[54] Kenta Oku, Takashi Nishizaki, and Fumio Hattori. Extraction of region-restricted phrases from geotagged contents based on region-restrictedness score. *Journal of IPSJ:Transactions on Databases (TOD)*, 5(3):97–116, 2012.

[55] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999.

[56] Marco Pennacchiotti and Ana-Maria Popescu. Democrats, republicans and starbucks afficionados: User classification in twitter. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 430–438, New York, NY, USA, 2011. ACM.

[57] Ashwin Rajadesingan, Reza Zafarani, and Huan Liu. Sarcasm detection on twitter: A behavioral modeling approach. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, pages 97–106, New York, NY, USA, 2015. ACM.

[58] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[59] Zhaochun Ren, Maria-Hendrike Peetz, Shangsong Liang, Willemijn van Dolen, and Maarten de Rijke. Hierarchical multi-label classification of social text streams. In *Proceedings of the 37th International ACM SIGIR Conference on*

Research & Development in Information Retrieval, SIGIR '14, pages 213–222, New York, NY, USA, 2014. ACM.

[60] Martin Riedl and Chris Biemann. Topictiling: A text segmentation algorithm based on lda. In *Proceedings of ACL 2012 Student Research Workshop*, ACL '12, pages 37–42, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[61] T. Sakaki, Y. Matsuo, T. Yanagihara, Naiwala P. Chandrasiri, and K. Nawa. Real-time event extraction for driving information from social sensors. In *Proceedings of the 2012 IEEE International Conference on Cyber Technology in Automation, Control and Intelligent Systems*, pages 221–226, 2012.

[62] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 851–860, New York, NY, USA, 2010. ACM.

[63] Claude E. Shannon. Prediction and entropy of printed english. *The Bell System Technical Journal*, 30:50–62, 1951.

[64] Bharath Sriram, Dave Fuhry, Engin Demir, Hakan Ferhatosmanoglu, and Murat Demirbas. Short text classification in twitter to improve information filtering. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 841–842, New York, NY, USA, 2010. ACM.

[65] V. Suresh, Avanthi Krishnamurthy, Rama Badrinath, and C.E. Veni Madhavan. *Advances in Intelligent Data Analysis X*, volume 7014, pages 364–375. Springer, 2011.

[66] Tetsuro Takahashi and Yuya Noda. Can twitter be an alternative of real-world sensors? *IEICE Technical Report*, 110(400):43–48, 2011 (in Japanese).

[67] Taiki Takano and Ushio Inoue. An extraction method of experience information from blogs based on sentence strucutre. In *The 3rd Forum on Data Engineering and Information Management*, A4-2, 2011 (in Japanese).

[68] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[69] Nguyen Minh The, Takahiro Kawamura, Hiroyuki Nakagawa, Yasuyuki Tahara, and Akihiko Ohsuga. Automatic mining of human activity attributes from weblogs. In *Proceedings of the IEEE/ACIS 9th International Conference on Computer and Information Society*, pages 633–638, 2010.

[70] Kei Tsuchiya, Masashi Toyoda, and Masaru Kitsuregawa. Extracting details of train troubles from microblogs. *IEICE Technical Report*, 2013-IFAT-111(31):1–6, 2013 (in Japanese).

[71] Twitter. Twitter reports fourth quarter and fiscal year 2013 results. https://investor.twitterinc.com/releasedetail.cfm?ReleaseID=823321, Feb. 2014.

[72] E. Voorhees and D. Tice. The trec-8 question answering track evaluation. In *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*, pages 77–82, 1999.

[73] Beidou Wang, Can Wang, Jiajun Bu, Chun Chen, Wei Vivian Zhang, Deng Cai, and Xiaofei He. Whom to mention: Expand the diffusion of tweets by @ recommendation on micro-blogging systems. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 1331–1340, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.

[74] Chi Wang, Marina Danilevsky, Nihit Desai, Yinan Zhang, Phuong Nguyen, Thrivikrama Taula, and Jiawei Han. A phrase mining framework for recursive construction of a topical hierarchy. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 437–445, New York, NY, USA, 2013. ACM.

[75] Xuanhui Wang, ChengXiang Zhai, Xiao Hu, and Richard Sproat. Mining correlated bursty topic patterns from coordinated text streams. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '07, pages 784–793, 2007.

[76] Zhihua Wei, Hongyun Zhang, Zhifei Zhang, Wen Li, and Duoqian Miao. A naive bayesian multi-label classification algorithm with application to visualize text search results. *International Journal of Advanced Intelligence*, 3(2):173–188, 2011.

[77] B. L. Welch. The generalization of 'student's' problem when several different population variances are involved. *Biometrika*, 34(1/2):28–35, 1947.

[78] Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. Twitterrank: Finding topic-sensitive influential twitterers. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 261–270, 2010.

[79] Ting-Fan Wu, Chih-Jen Lin, and Ruby C. Weng. Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, 5:975–1005, 2004.

[80] Wei Wu, Bin Zhang, and Mari Ostendorf. Automatic generation of personalized annotation tags for twitter users. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 689–692, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[81] Wei Xie, Feida Zhu, Jing Jiang, Ee-Ping Lim, and Ke Wang. Topicsketch: Real-time bursty topic detection from twitter. In *Proceedings of the 2013 IEEE 13th International Conference on Data Mining*, pages 837–846, 2013.

[82] Yuto Yamaguchi, Toshiyuki Amagasa, and Hiroyuki Kitagawa. Tagging users based on twitter lists. *International Journal Web Engineering Technology*, 7(3):273–298, aug 2012.

[83] Yuto Yamaguchi, Toshiyuki Amagasa, Hiroyuki Kitagawa, and Yohei Ikawa. Online user location inference exploiting spatiotemporal correlations in social streams. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 1139–1148, New York, NY, USA, 2014. ACM.

[84] Masahito Yamamoto, Hiroya Ogasawara, Ikuo Suzuki, and Masashi Furukawa. Tourism informatics:9. information propagation network for 2012 tohoku earthquake and tsunami on twitter. *IPSJ Magazine*, 53(11):1184–1191, 2012 (in Japanese).

[85] Shuhei Yamamoto and Tetsuji Satoh. Two phase estimation method for multi-classifying real life tweets. *International Journal of Web Information Systems*, 10(4):378–393, 2014.

[86] Yuan Cao Zhang, Diarmuid Ó Séaghdha, Daniele Quercia, and Tamas Jambor. Auralist: Introducing serendipity into music recommendation. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, pages 13–22, New York, NY, USA, 2012. ACM.

[87] Wayne Xin Zhao, Jing Jiang, Jing He, Yang Song, Palakorn Achananuparp, Ee-Peng Lim, and Xiaoming Li. Topical keyphrase extraction from twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational*

Linguistics: Human Language Technologies, HLT '11, pages 379–388, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[88] Zhe Zhao and Qiaozhu Mei. Questions about questions: An empirical analysis of information needs on twitter. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 1545–1556, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.

[89] Xingwei Zhu, Zhao-Yan Ming, Yu Hao, Xiaoyan Zhu, and Tat-Seng Chua. Customized organization of social media contents using focused topic hierarchy. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 1509–1518, New York, NY, USA, 2014. ACM.

Publication List

Reference Publications

Journal

1. Shuhei Yamamoto and Tetsuji Satoh, "A Study on Upgrading Precision of Aspect Estimation for Real Life Tweets," IPSJ Journal, Vol. 56, No. 6, pp. 1496-1506, June 2015 (in Japanese with English Abstract).

2. Shuhei Yamamoto and Tetsuji Satoh, "Two Phase Estimation Method for Multi-classifying of Real Life Tweets," International Journal of Web Information Systems, Vol. 10, Issue 4, pp. 343-362, October 2014.

3. Shuhei Yamamoto and Tetsuji Satoh, "Multi-label Classification for Real Life Tweets Based on Association between Topics and Aspects," IPSJ Transactions on Databases (TOD), Vol. 7, No. 2, pp. 24-36, June 2014 (in Japanese with English Abstract).

Proceedings

1. Shuhei Yamamoto, Noriko Kando, and Tetsuji Satoh, "LAIM: Life Aspect Inference Method based on Probability Distribution for Real Life Tweets," The 2015 IEEE/WIC/ACM International Conference on Web Intelligence (WI2015), Singapore, pp. 187-194, December 2015.

2. Shuhei Yamamoto and Tetsuji Satoh, "Hierarchical Estimation Framework of Multi-label Classifying: A Case of Tweets Classifying into Real Life Aspects," The 15th International AAAI Conference on Web and Social Media (ICWSM2015), Oxford, UK, pp. 523-532, May 2015.

3. Shuhei Yamamoto and Tetsuji Satoh, "Two Phase Extraction Method for Multi-label Classification of Real Life Tweets," The 15th International Conference on Information Integration and Web-based Applications & Services (iiWAS2013), Vienna, Austria, pp. 16-25, December 2013.

4. Shuhei Yamamoto and Tetsuji Satoh, "Two Phase Extraction Method for Extracting Real Life Tweets using LDA," The 15th Asia-Pacific Web Conference (APWeb2013), Sydney, Australia, pp. 340-347, April 2013.

Other Publications

Journal

1. Yutaro Yamaguchi, Shuhei Yamamoto, and Tetsuji Satoh, "Behavior Analysis Methods for Twitter Users based on Transitions in Posting Activities," International Journal of Web Information Systems, Vol. 10, Issue 4, pp. 363-377, October 2014.

2. Yuhiro Mizunuma, Atsushi Ikeuchi, Shuhei Yamamoto, Yutaro Yamaguchi, Tetsuji Satoh, and Satoshi Shimada, "Analysis of the Occurrence Factor and Classification of Bursty Status on Twitter of Infosociomics," The Journal of Infosocionomics Society, Vol. 7, No. 2, pp. 41-50, March 2013 (in Japanese with English Abstract).

Proceedings

1. Shuhei Yamamoto, Kei Wakabayashi, Noriko Kando, and Tetsuji Satoh, "BUTE: Bursty Users Tagging Method Estimated by Time Series Data," The 17th International Conference on Information Integration and Web-based Applications & Services (iiWAS2015), Brussels, Belgium, pp. 148-156, December 2015.

2. Hideo Joho, Adam Jatowt, Roi Blanco, Hajime Naka, and Shuhei Yamamoto, "Overview of NTCIR-11 Temporal Information Access (Temporalia) Task," NII Testbeds and Community for Informaiton access Research (NTCIR-11) Conference, Tokyo, Japan, pp. 429-437, December 2014.

3. Yuhiro Mizunuma, Shuhei Yamamoto, Yutaro Yamaguchi, Atsushi Ikeuchi, Tetsuji Satoh, and Satoshi Shimada, "Twitter Bursts: Analysis of their Occurrences and Classifications," The 8th International Conference on Digital Society (ICDS2014), Barcelona, Spain, pp. 182-187, March 2014.

4. Yutaro Yamaguchi, Shuhei Yamamoto, and Tetsuji Satoh, "Behavior Analysis of Microblog Users Based on Transitions in Posting Activities," The 15th International Conference on Information Integration and Web-based Applications & Services (iiWAS2013), Vienna, Austria, pp. 63-67, December 2013.